Study of Subjective and Objective Quality Assessment of Video

Kalpana Seshadrinathan, *Member, IEEE*, Rajiv Soundararajan, *Student Member, IEEE*, Alan Conrad Bovik, *Fellow, IEEE*, and Lawrence K. Cormack

***Abstract*—We present the results of a recent large-scale subjective study of video quality on a collection of videos distorted by a variety of application-relevant processes. Methods to assess the visual quality of digital videos as perceived by human observers are becoming increasingly important, due to the large number of applications that target humans as the end users of video. Owing to the many approaches to video quality assessment (VQA) that are being developed, there is a need for a diverse independent public database of distorted videos and subjective scores that is freely available. The resulting Laboratory for Image and Video Engineering (LIVE) Video Quality Database contains 150 distorted videos (obtained from ten uncompressed reference videos of natural scenes) that were created using four different commonly encountered distortion types. Each video was assessed by 38 human subjects, and the difference mean opinion scores (DMOS) were recorded. We also evaluated the performance of several state-of-the-art, publicly available full-reference VQA algorithms on the new database. A statistical evaluation of the relative performance of these algorithms is also presented. The database has a dedicated web presence that will be maintained as long as it remains relevant and the data is available online.**

***Index Terms*—Full reference, human visual system, LIVE video quality database, perceptual quality assessment, video quality, visual perception.**

I. Introduction

DIGITAL videos are increasingly finding their way into the day-to-day lives of people via the explosion of video applications such as digital television, digital cinema, Internet videos, video teleconferencing, video-sharing services such as Youtube, Video On Demand (VoD), home videos, and so on. Digital videos typically pass through several processing stages before they reach the end user of the video. Most often, this end user is a human observer. The effect of most processing stages is to degrade the quality of the video that passes through it, although certain processing stages (for example, in consumer devices) attempt to improve quality. Methods for evaluating video quality play a critical role in quality monitoring to maintain Quality of Service (QoS) requirements; performance evaluation of video acquisition and display devices; evaluation of video processing systems for compression, enhancement, error concealment, and so on; and finally, perceptually optimal design of video processing systems.

The only reliable method to assess the video quality perceived by a human observer is to ask human subjects for their opinion, which is termed subjective video quality assessment (VQA). Subjective VQA is impractical for most applications due to the human involvement in the process. However, subjective VQA studies provide valuable data to assess the performance of *objective* or automatic methods of quality assessment. In addition to providing the means to evaluate the performance of state-of-the-art VQA technologies, subjective studies also enable improvements in the performance of VQA algorithms toward attaining the ultimate goal of matching human perception.

In this paper, we first present a study that we conducted to assess the subjective quality of videos. Our study included 10 uncompressed reference videos of natural scenes and 150 distorted videos (obtained from the references) using four different distortion types commonly encountered in applications. Each video was assessed by 38 human subjects in a single stimulus study with hidden reference removal, where the subjects scored the video quality on a continuous quality scale. This study and the resulting video database presented here, which we call the Laboratory for Image and Video Engineering (LIVE) Video Quality Database, supplements the widely used LIVE Image Quality Database for still images [1]. We evaluate the performance of leading, publicly available objective VQA algorithms on the new LIVE Video Quality Database by using standardized measures. This paper builds upon our earlier work describing the LIVE Video Quality Database [2].

Currently, the only publicly available subjective data that is widely used by the VQA community comes from the study conducted by the Video Quality Experts Group (VQEG) as part of its FR-TV Phase 1 project in 2000 [3]. There have been significant advances in video processing technology since 2000, most notably the development of the popular H.264/MPEG-4 AVC compression standard. The test videos in the VQEG study are not representative of present generation encoders and communication systems. By contrast, the LIVE Video Quality Database described here includes videos distorted by H.264 compression, as well as videos resulting from simulated transmission of H.264 packetized streams through error prone communication channels. The VQEG study targeted secondary

Manuscript received April 15, 2009; revised December 28, 2009. First published February 02, 2010; current version published May 14, 2010. This research was supported by a grant from the National Science Foundation under Award 0728748. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Eero P. Simoncelli.

K. Seshadrinathan is with Intel Corporation, Chandler, AZ 85226 USA (e-mail: Kalpana.Seshadrinathan@ieee.org).

R. Soundararajan is with the University of Texas at Austin, Austin, TX 78712 USA (e-mail: rajiv.s@mail.utexas.edu).

A. C. Bovik is with the Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: bovik@ece.utexas.edu).

L. K. Cormack is with the Psychology Department, University of Texas at Austin, Austin, TX 78712 USA (e-mail: cormack@psy.utexas.edu).

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/TIP.2010.2042111

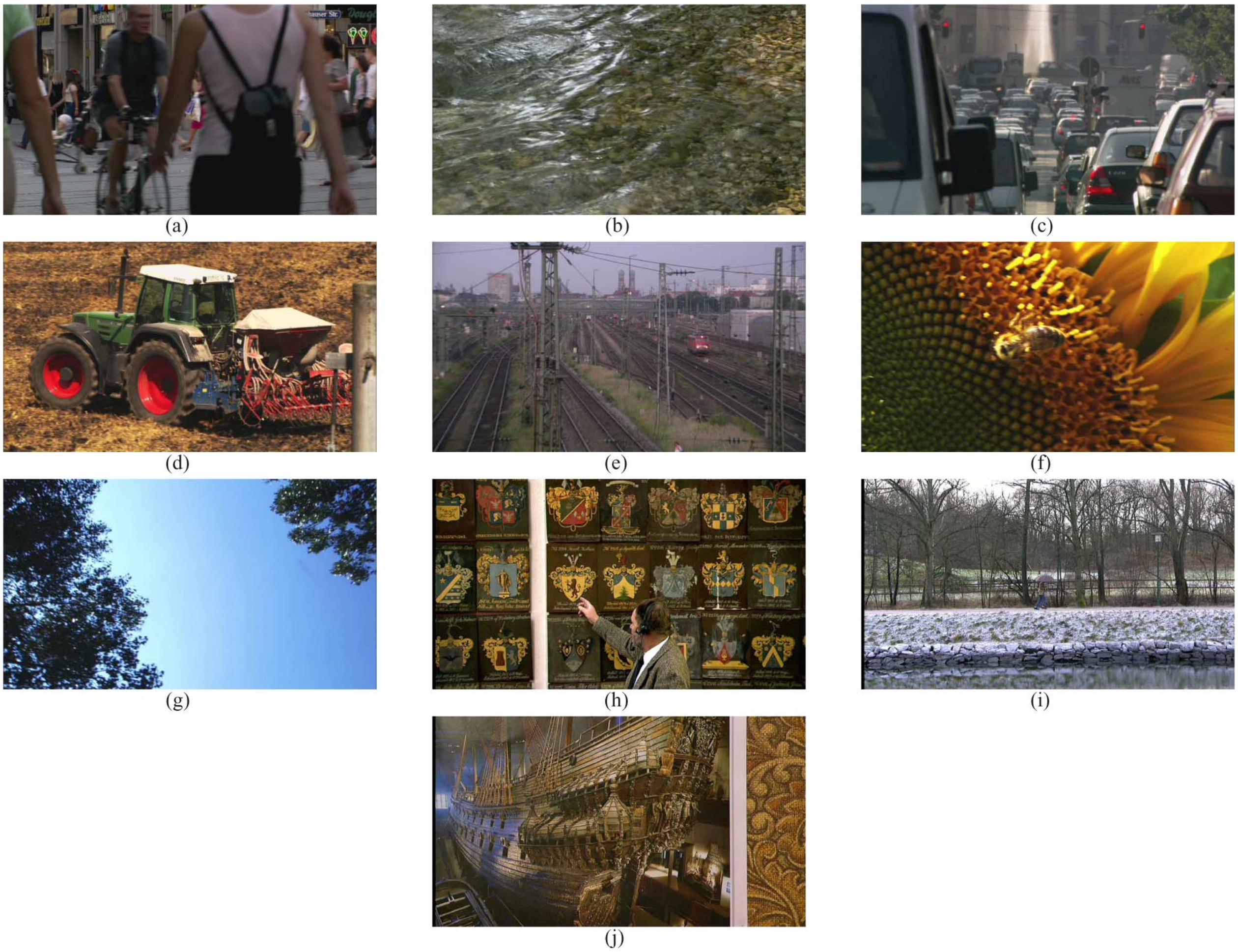


Fig. 1. One frame from each of the ten reference videos used in the study. (a) Pedestrian Area. (b) River Bed. (c) Rush Hour. (d) Tractor. (e) Station. (f) Sunflower. (g) Blue Sky. (h) Shield (i) Park Run. (j) Mobile & Calendar.

distribution of television, so most of the videos in the VQEG study are interlaced. Interlaced videos lead to visual artifacts in the reference as well as distorted videos when they are displayed in increasingly common progressive scan monitors. Objective VQA algorithms typically involve multiple processing steps which require adjustment to handle interlaced signals. De-interlacing creates visual artifacts associated with the particular algorithm used, which is unacceptable in a VQA framework. Additionally, interlaced videos are not representative of current trends in the video industry such as multimedia, IPTV, video viewing on computer monitors, progressive High Definition Television (HDTV) standards, and so on. Videos in the LIVE Video Quality Database were all captured in progressive scan formats, allowing researchers to focus on developing algorithms for VQA. Further, the VQEG database was designed to address the needs of secondary distribution of television and hence, the database spans narrow ranges of quality scores—indeed, more than half of the sequences are of very high quality (MPEG-2 encoded at > 3 Mbps). Overall, the VQEG videos exhibit poor perceptual separation, making it difficult to distinguish the performance of VQA algorithms. The LIVE Video Quality Database spans a much wider range of quality—the low-quality videos were designed to be of similar quality found in streaming video applications on the Internet (Youtube, wireless videos, live streaming of low-bandwidth videos, etc.).

Although the VQEG has several other completed and ongoing projects, none of the videos from subsequent studies have been made public [4], [5]. Only subjective data has been made available publicly from the VQEG FRTV Phase 2 study and the videos have not been made public, due to several copyright and licensing issues [6]. The situation with the VQEG Multimedia dataset is identical, wherein the VQEG plans to release only the subjective data in September, 2009 and the videos will not be released publicly [7]. This is a grave concern, since unavailability of the VQEG datasets seriously limits the ability of researchers to benchmark the performance of new, objective VQA models against the VQEG evaluations. The LIVE Video Quality Database is publicly available for download from [8] to facilitate comparative evaluation of newer objective models and to advance the state-of-the-art in perceptual quality evaluation systems.

II. Details of Subjective Study

A. *Source Sequences*

We used ten uncompressed, high-quality, source videos of natural scenes (as opposed to animation, graphics, text, etc.) that

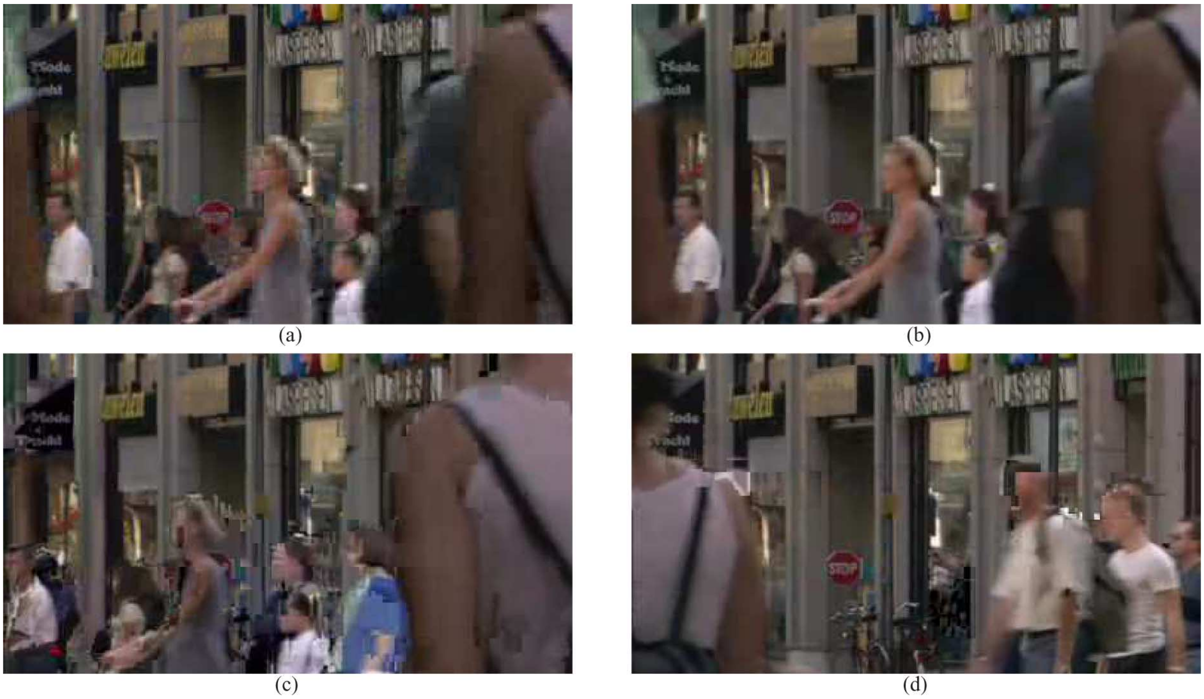

(a) (b) (c) (d)

Fig. 2. (a) MPEG-2 compressed frame (b) H.264 compressed frame (c) IP loss simulated frame (d) Wireless loss simulated frame.

are freely available for download from the Technical University of Munich [9]. All videos provided by [9] were filmed with professional, high-end equipment and converted to digital format with utmost care, guaranteeing that the reference videos are distortion free. We only used the progressively scanned videos in this database, thus avoiding problems with video deinterlacing. We used the digital videos provided in High Definition (HD) YUV 4:2:0 format and none of the videos contain audio components. However, due to resource limitations when displaying these videos, we downsampled all videos to a resolution of 768 × 432 pixels. We chose this resolution to ensure that the aspect ratio of the HD videos was maintained, thus minimizing visual distortions. Additionally, this resolution ensures that the number of rows and columns are multiples of 16, as is often required by compression systems such as MPEG-2. We downsampled each raw video frame by frame using the "imresize" function in Matlab using bicubic interpolation to minimize distortions due to aliasing.

Fig. 1 shows one frame of each reference video in the LIVE Video Quality Database. All videos, except blue sky, are 10 s long. The blue sky sequence is 8.68 s long. The first seven sequences have a frame rate of 25 frames per second, while the remaining three (Park run, Shields, and Mobile & Calendar) have a frame rate of 50 frames per second. A short description of these videos is provided below.

- *Blue Sky*—Circular camera motion showing a blue sky and some trees
- *River Bed*—Still camera, shows a river bed containing some pebbles and water
- *Pedestrian area*—Still camera, shows some people walking about in a street intersection
- *Tractor*—Camera pan, shows a tractor moving across some fields
- *Sunflower*—Still camera, shows a bee moving over a sunflower in close-up
- *Rush hour*—Still camera, shows rush hour traffic on a street
- *Station*—Still camera, shows a railway track, a train, and some people walking across the track
- *Park run*—Camera pan, a person running across a park
- *Shields*—Camera pans at first, then becomes still and zooms in; shows a person walking across a display pointing at it
- *Mobile & Calendar*—Camera pan, toy train moving horizontally with a calendar moving vertically in the background

B. Test Sequences

We created 15 test sequences from each of the reference sequences using four different distortion processes—MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bitstreams through error-prone IP networks, and through error-prone wireless networks. The goal of our study was to develop a database of videos that will challenge automatic VQA algorithms. We included diverse distortion types to test the ability of objective models to predict visual quality consistently across distortions. Compression systems such as MPEG-2 and H.264 produce fairly uniform distortions/quality in the video, both spatially and temporally. Network losses, however, cause *transient* distortions in the video, both spatially and temporally. Fig. 2 shows part of a frame of the "Pedestrian Area" sequence corrupted by each of the four distortion types in the LIVE Video Quality Database. It is clear that the visual appearance of distortion is very different in each of these videos. MPEG-2 and H.264 compressed videos exhibit typical compression artifacts such as blocking, blur, ringing and motion compensation mismatches around object edges. Notice, however, the difference in the distortions created by the MPEG-2 and H.264 compression systems, such as reduced blockiness in the H.264 compressed frame. Videos obtained from lossy transmission through wireless networks exhibit errors that are restricted to small regions of a frame. Videos obtained from lossy transmission through IP networks exhibit errors in larger regions of the frame. Errors in wireless and IP networks are also *temporally transient* and appear as glitches in the video. Almost half the videos in the LIVE Video Quality Database contain spatio-temporally localized distortions, while the VQEG Phase 1 dataset is largely comprised of compressed videos and contains only a few videos with errors and spatio-temporally localized distortions.

The distortion strengths were adjusted manually so that the videos obtained from each source and each distortion category spanned a set of contours of equal visual quality. A large set of videos were generated and viewed by the authors and a subset of these videos that spanned the desired visual quality were chosen to be included in the LIVE Video Quality Database. To illustrate this procedure, consider four labels for visual quality ("Excellent," "Good," "Fair," and "Poor") and one reference video ("Tractor"). Four MPEG-2 compressed versions of "Tractor" are chosen to approximately match the four labels for visual quality. Similar procedure is applied to select H.264 compressed, wireless and IP distorted versions of "Tractor." Note that the "Excellent" MPEG-2 video and "Excellent" H.264 video are designed to have the approximate same visual quality and similarly for other distortion categories and quality labels. The same selection procedure is then repeated for every reference video. Note that an "Excellent" test video obtained from "Sunflower" is designed to have the approximate same visual quality as an "Excellent" test video obtained from "Tractor" and similarly for other reference videos. Our design of the distorted videos tests the ability of objective VQA models to predict visual quality consistently across varying content and distortion types. The LIVE Video Quality Database is unique in this respect and we believe that adjusting distortion strength perceptually, as we have done here, is far more effective toward challenging and distinguishing the performance of objective VQA algorithms than, for instance, fixing the compression rates across sequences as is done in most studies including the VQEG FR-TV Phase 1 study [3]. The four distortion types are detailed in Sections II-B1–B4.

1) MPEG-2 Compression: The MPEG-2 standard is used in a wide variety of video applications, most notably DVD's and digital broadcast television. There are four MPEG-2 compressed videos corresponding to each reference in our database and we will refer to this distortion category as "MPEG-2" in the remainder of the paper. We used the MPEG-2 reference software available from the International Organization for Standardization (ISO) to compress the videos [10].

The bit rate required to compress videos for a specified visual quality varies dramatically depending on the content. The authors selected four compressed MPEG-2 videos for each reference video by viewing compressed videos generated using a wide variety of bit rates and selecting a subset that spanned the desired range of visual quality. "Excellent" quality videos were chosen to be quite close to the reference in visual quality. "Poor" quality videos were chosen to be of similar quality as Youtube videos, without being obliterated by MPEG blocking artifacts. The compression rates varied from 700 kbps to 4 Mbps, depending on the reference sequence.

2) H.264 Compression: H.264 is rapidly gaining popularity due to its superior compression efficiency as compared to MPEG-2. There are four H.264 compressed videos corresponding to each reference in our database and we will refer to this distortion category as "H.264" in the remainder of the paper. We used the JM reference software (Version 12.3) made available by the Joint Video Team (JVT) [11].

The procedure for selecting the videos was the same as that used to select MPEG-2 compressed videos. The compression rates varied from 200 kbps to 5 Mbps.

3) Transmission Over IP Networks: Videos are often transmitted over IP networks in applications such as video telephony and conferencing, IPTV and Video on Demand. There are three "IP" videos corresponding to each reference in our database that were created by simulating IP losses on an H.264 compressed video stream and we will refer to this distortion category as "IP" in the remainder of the paper. The H.264 compressed video streams were created using the JM reference software [11] and compression rates varied between 0.5–7 Mbps.

An in-depth study of the transport of H.264 video over IP networks can be found in [12] and many of our design considerations in the video communication system were based on this study. IP networks offer best effort service and packet losses occur primarily due to buffer overflow-at intermediate nodes in a network with congestion. The video sequences subjected to errors in the IP environment contained between one and four slices per frame and each packet contained one slice; we only used these two options since they result in packet sizes that are typical in IP networks. Using one slice per frame has the advantage of reducing overhead due to IP headers, but at the expense of robustness [12]. Using four slices per frame increases robustness to error (likelihood of an entire frame getting lost is reduced), at the expense of reducing compression efficiency.

Four IP error patterns supplied by the Video Coding Experts Group (VCEG), with loss rates of 3%, 5%, 10%, and 20%, were used [13]. The error patterns were obtained from real-world experiments on congested networks and are recommended by the VCEG to simulate the Internet backbone performance for video coding experiments. We created test videos by dropping packets specified in the error pattern from an H.264 compressed packetized video stream. To enable decoding, we did not drop the first packet [containing the Instantaneous Data Refresh (IDR)] and the last packet (since the loss of this packet cannot be detected by the decoder). This is equivalent to assuming that these packets were transmitted reliably out of band. The resulting H.264 bitstream was then decoded using [11] and the losses concealed using the built-in error concealment mechanism (mode 2—motion copy) [14].

The authors viewed and selected a diverse set of videos suffering from different types of observed artifacts and spanning the desired range of quality. The type of observed artifact varies depending on the following:

- *Whether an Intracoded frame (I frame) or Predicted frame (P frame) is lost*—I frame losses result in much more severe and sustained video distortions (that last until the next I frame is received correctly).
- *Whether each frame is transmitted in 1 packet or 4 packets*—Loss of an entire frame when transmitted as a

single slice results in much more significant distortions, than when the frame is transmitted using four slices.

- *Flexible Macroblock Ordering (FMO)*—We used both regular and dispersed modes of FMO in our simulations [15]. In dispersed mode, we used four packet groups formed by subsampling the frame by 2 along both rows and columns. Loss of video packets in regular mode results in severe artifacts in localized regions of the video, while the impairments are not as severe in the dispersed mode.

4) Transmission Over Wireless Networks: Video transmission for mobile terminals is envisioned to be a major application in 3G systems and the superior compression efficiency and error resilience of H.264 makes it ideal for use in harsh wireless transmission environments [15]. There are four videos corresponding to each reference in our database that were created by simulating losses sustained by an H.264 compressed video stream in a wireless environment and we will refer to this distortion category as "Wireless" in the remainder of the paper. The H.264 compressed bitstreams were created using the JM reference software [11] and compression rates varied between 0.5–7 Mbps.

An in-depth study of the transport of H.264 video over wireless networks can be found in [15]. Many of our design considerations for the wireless simulations was based on this study. A packet transmitted over a wireless channel is susceptible to bit errors due to attenuation, shadowing, fading and multiuser interference in wireless channels. We assume that a packet is lost even if it contained a single bit error, an assumption that is often made in practice [15]. Due to this assumption, a longer packet is more likely to be lost and shorter packet sizes are desirable in wireless networks. We encoded the video stream using multiple slices per frame, where each packet contained one slice. All packets contained roughly the same number of bytes (approximately 200 bytes per packet), making their susceptibility to bit errors almost identical. We simulated errors in wireless environments using bit error patterns and software available from the VCEG [16]. The packet error rates using these bit error patterns varied between 0.5–10%. The decoding and error concealment techniques for the wireless simulations were identical to the IP simulations.

Again, the authors viewed and selected videos suffering from different types of observed artifacts and spanning the desired range of quality. Observed artifacts in the wireless environment also depend on whether an I or P packet is lost and on the FMO mode. Due to the smaller packet sizes in wireless applications, the observed artifacts are spatio-temporally localized and appear different from the artifacts observed in IP applications.

C. Subjective Testing Design

We adopted a single stimulus continuous procedure to obtain subjective quality ratings for the different video sequences. The choice of a single stimulus paradigm is well suited to a large number of emerging multimedia applications, such as quality monitoring for Video on Demand, IPTV, Internet streaming, etc. Additionally, it significantly reduces the amount of time needed to conduct the study (given a fixed number of human subjects) as compared to a double stimulus study. The subjects indicated the quality of the video on a continuous scale. The continuous scale allows the subject to indicate fine gradations in visual quality. We believe this is superior to the ITU-R Absolute Category Rating (ACR) scale that uses a five-category quality judgment, as is used in recent VQEG studies [5]. The subject also viewed each of the reference videos to facilitate computation of Difference Mean Opinion Scores (DMOS), a procedure known as hidden reference removal [17], [18].

All the videos in our study were viewed by each subject, which required half an hour of the subject's time. To minimize the effects of viewer fatigue, we conducted the study in two sessions of 30 minutes each.

We prepared playlists for each subject by arranging the 150 test videos in a random order using a random number generator. We did not want the subjects to view successive presentations of test videos that were obtained from the same reference sequence, to avoid contextual and memory effects in their judgment of quality. Once a playlist was constructed, adjacent sequences were examined to determine if they corresponded to the same content. If any such pairs were detected, one of the videos was swapped with another randomly chosen video in the playlist which did not suffer from the same problem. This list was then split into two halves for the two sessions.

We wanted to ensure that any differences in the use of the quality judgment scale by the subject between sessions did not affect the results of the study. For instance, a subject may be very critical of the visual quality of a video in one session and more forgiving in the other. To avoid this problem, we included each reference video in both sessions in the hidden reference removal process. We inserted each of the ten reference videos into the playlists for each session randomly, again ensuring that successive playback of the same content did not occur. The DMOS scores were then computed for each video *per session* using the quality score assigned to the reference video in that session, as described in Section III.

D. Subjective Testing Display

We developed the user interface for the study on a Windows PC using MATLAB, in conjunction with the XGL toolbox for MATLAB developed at the University of Texas at Austin [19]. The XGL toolbox allows precise presentation of psychophysical stimuli to human observers. It is extremely important to avoid any errors in displaying the video such as latencies or frame drops. This can significantly affect the results of the study since the subject's quality perception is affected not by the video itself, but by the display issues. To ensure perfect playback, all distorted sequences were processed and stored as raw YUV 4:2:0 files. An entire video was loaded into memory before its presentation began to avoid any latencies due to slow hard disk access of large video files. The videos were then played out at the appropriate frame rate for the subject to view. The XGL toolbox interfaces with the ATI Radeon X600 graphics card in the PC and utilizes its ability to play out the YUV videos. The videos were viewed by the subjects on a cathode ray tube (CRT) monitor to avoid the effects of motion blur and low refresh rates on liquid crystal display (LCD) monitors. The entire study was conducted using the same monitor and we calibrated the CRT monitor using the Monaco Optix XR Pro device. The XGL toolbox avoids visual artifacts by synchronizing the display so that the switching between adjacent frames of the video occurs during the retrace of the CRT scan. Since the videos had low frame rates (25 and 50 Hz), we set the monitor resolution to 100 Hz to avoid artifacts due to monitor flicker. Each frame of

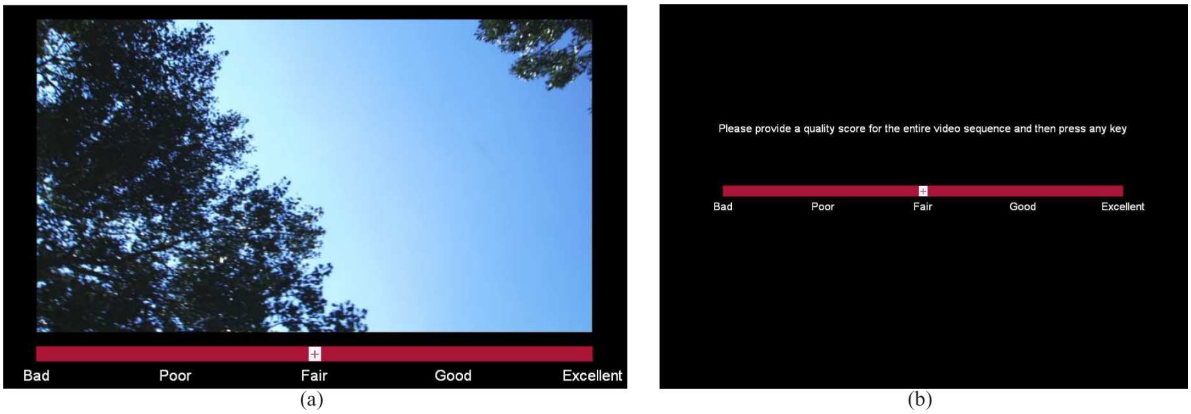

Please provide a quality score for the entire video sequence and then press any key

Fig. 3. (a) Screenshot from the subjective study interface displaying the video to the subject. (b) Screenshot from the subjective study interface that prompts the subject to enter a quality score for the video they completed viewing.

the 50-Hz videos was displayed for two monitor refresh cycles and each frame of the 25-Hz videos was displayed for four monitor refresh cycles.

The screen was set at a resolution of 1024 × 768 pixels and the videos were displayed at their native resolution to prevent any distortions due to scaling operations performed by software or hardware. The remaining areas of the display were black. At the end of the presentation of the video, a continuous scale for video quality was displayed on the screen, with a cursor set at the center of the quality scale to avoid biasing the subject's quality percept. The quality scale had five labels marked on it to help the subject. The left end of the scale was marked "Bad" and the right end was marked "Excellent." Three equally spaced labels between these were marked "Poor," "Fair," and "Good," similar to the ITU-R ACR scale. Screenshots from the subjective study interface are shown in Fig. 3. The subject could move the cursor along the scale by moving a mouse. The subject was asked to press a key to enter the quality score after moving the cursor to a point on the scale that corresponded to his or her quality percept. The subject was allowed to take as much time as needed to enter the score. However, the subject could not change the score once entered or view the video again. Once the score was entered, the next video was displayed.

E. Subjects and Training

All subjects taking part in the study were recruited from the undergraduate Digital Image and Video Processing class (fall 2007) at the University of Texas at Austin. The subject pool consisted of mostly male students. The subjects were not tested for vision problems. Each video was ranked by 38 subjects.

Each subject was individually briefed about the goal of the experiment and viewed a short training session before starting the experiment. Subjects viewed six training videos in their first session of participation and three training videos in their second session. Subjects were asked to provide quality scores for the training videos also to familiarize themselves with the testing procedure. The training videos were not part of the database and contained different content. The training videos were of 10-s duration and were also impaired by the same distortions as the test videos. We selected the training videos to span the same range of quality as the test videos, to give the subject an idea of the quality of videos they would be viewing in the study and to enable suitable use of the quality scale by the subject.

III. Processing of Subjective Scores

Let s_{ijk} denote the score assigned by subject i to video j in session $k = \{1, 2\}$. Since our focus in this paper is on full-reference objective VQA algorithms that assume a "perfect" reference video, we compute difference scores between the test video and the corresponding reference to discount any subject preferences for certain reference videos. First, difference scores d_{ijk} are computed per session by subtracting the quality assigned by the subject to a video from the quality assigned by the same subject to the corresponding reference video *in the same session*. Computation of difference scores per sessions helps account for any variability in the use of the quality scale by the subject between sessions

$$d_{ijk} = s_{ij_{\mathrm{ref}}k} - s_{ijk}. \tag{1}$$

The difference scores for the reference videos are 0 in both sessions and are removed. The difference scores per session are then converted to Z-scores *per session* [20]

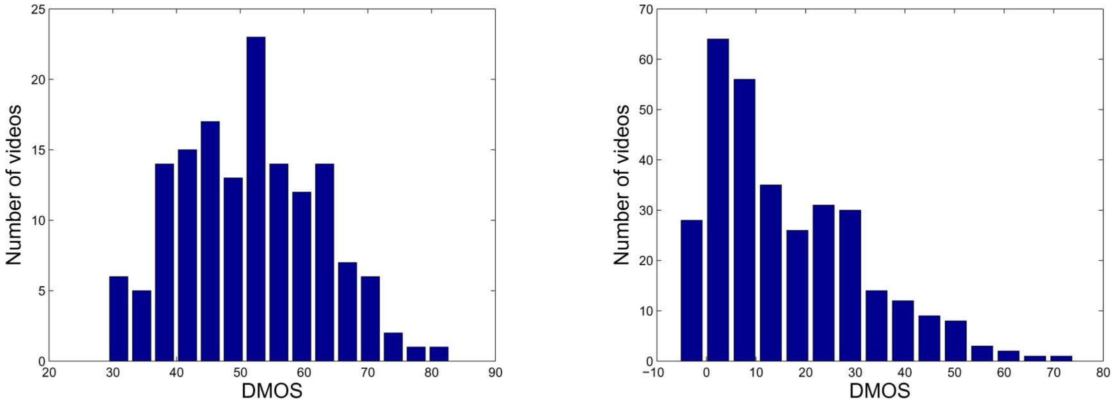
$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} d_{ijk} \tag{2}$$

$$\sigma_{ik} = \sqrt{\frac{1}{N_{ik}-1} \sum_{j=1}^{N_{ik}} (d_{ijk} - \mu_{ik})^2} \tag{3}$$

$$z_{ijk} = \frac{d_{ijk} - \mu_{ik}}{\sigma_{ik}} \tag{4}$$

where N_{ik} is the number of test videos seen by subject i in session k. Again, note that Z-scores are computed per session to account for any differences in the use of the quality scale (differences in the location and range of values used by the subject) between sessions.

Every subject sees each test video in the database exactly once, either in the first session or in the second session. The Z-scores from both sessions are then combined to create a matrix $\{z_{ij}\}$ corresponding to the Z-score assigned by subject i to



(a)

(b)

Fig. 4. Histogram of the DMOS scores in 15 equally spaced bins between the minimum and maximum DMOS values for (a) LIVE Video Quality Database and (b) VQEG FRTV Phase 1 Database.

video j, where $j = \{1, 2, \ldots, N\}$ indexes $N = 150$ test videos in the LIVE Video Quality Database.

A subject rejection procedure specified in the ITU-R BT 500.11 recommendation is then used to discard scores from unreliable subjects [21]. Note that Z-scores in (4) account for any differences in subject preferences for reference videos, use of the quality scale between subjects, and differences in use of the quality scale by a subject between sessions. We believe that the processing and subject rejection procedure used here is superior to the VQEG studies for these reasons [3], [6], [7]. The ITU-R BT 500.11 recommendation first determines if the scores assigned by a subject are normally distributed by computing the kurtosis of the scores. The scores are considered normally distributed if the kurtosis falls between the values of 2 and 4. If the scores are normally distributed, the procedure rejects a subject whenever more than 5% of scores assigned by him falls outside the range of two standard deviations from the mean scores. If the scores are not normally distributed, the subject is rejected whenever more than 5% of his scores falls outside the range of 4.47 standard deviations from the mean scores. In both situations, care is taken to ensure that subjects who are consistently pessimistic or optimistic in their quality judgments are not eliminated [21]. In our study, 9 out of the 38 subjects were rejected at this stage. We found that the reason for the large number of rejected subjects is the borderline reliability of four subjects. The 5% criterion used in the subject rejection procedure translates to 7.5 videos in the LIVE Video Quality Database. Four of the nine rejected subjects scored 8 videos outside the expected range in the LIVE study and were rejected by the procedure.

Z-scores were then linearly rescaled to lie in the range of [0,100]. Assuming that Z-scores assigned by a subject are distributed as a standard Gaussian, 99% of the scores will lie in the range [-3,3] and we found that all Z-scores in our study fell inside this range. Rescaling was hence accomplished by linearly mapping the range [−3,3] to [0,100] using

$$z'_{ij} = \frac{100(z_{ij} + 3)}{6}. \tag{5}$$

Finally, the Difference Mean Opinion Score (DMOS) of each video was computed as the mean of the rescaled Z-scores from the $M = 29$ remaining subjects after subject rejection.

$$\text{DMOS}_j = \frac{1}{M} \sum_{i=1}^{M} z'_{ij}. \tag{6}$$

The LIVE Video Quality Database was designed to sample a range of visual quality in an approximately uniform fashion, as described in Section II-B. To illustrate this, we show histograms of the DMOS scores obtained from the LIVE Video Quality Database and the VQEG FRTV Phase 1 database in Fig. 4. Fig. 4 shows that the LIVE Video Quality Database exhibits reasonably uniform distribution of scores along the DMOS axis, while the VQEG FRTV Phase 1 database exhibits poor perceptual separation with a large number of videos of very high quality and far fewer videos of poor quality.

The DMOS scores in the LIVE Video Quality Database lie in the range $[30, 82]$, as seen in Fig. 4. This range may appear small to readers used to seeing subjective scores obtained using the highly popular Double Stimulus Continuous Quality Scoring (DSCQS) paradigm for subjective testing [21]. The DSCQS method was also used in the VQEG Phase 1 study, where the subjects score the quality of the reference and test videos on a [0,100] scale and DMOS is computed as the difference between the scores assigned to the reference and test video. The LIVE Video Quality Database, on the other hand, uses a single stimulus paradigm with hidden reference removal and DMOS is computed as *Z-scores* assigned by subjects, and not as differences between scores assigned to the reference and test videos. We believe that conversion of difference scores to Z-scores, as we have done here, is very important to account for differences in use of the scale by subjects. Assuming that Z-scores assigned by a subject are distributed as a standard Gaussian, 99% of Z-scores will lie in the range $[-3, 3]$ that corresponds to DMOS scores in the range $[0{,}100]$. $[30, 82]$ on the DMOS scale used in the LIVE Video Quality Database corresponds to mean Z-scores in the range $[-1.2, 1.92]$, which corresponds to approximately 86% of the area of the standard normal distribution. We believe that this range is reasonable for mean Z-scores, with individual Z-scores fluctuating beyond this range to extreme points on the scale.

IV. Objective VQA Algorithms

The performance of several publicly available objective VQA algorithms was evaluated on the LIVE Video Quality Database. One of the problems we faced was the lack of free availability of many VQA algorithms, since many popular VQA algorithms and tools are licensed and sold for profit. These include the Picture Quality Analyzer from Tektronix [22]; the Perceptual Evaluation of Video Quality (PEVQ) from Opticom [23]; the V-Factor from Symmetricom [24]; VQA solutions from SwissQual [25] and Kwill Corporation [26] and several others [27]. Our testing was limited to freely available VQA algorithms. Naturally, we will broaden our test set as more algorithms become freely available.

We tested the following VQA algorithms on the LIVE Video Quality Database.

- *Peak Signal-to-Noise Ratio (PSNR)* is a simple function of the Mean Squared Error (MSE) between the reference and test videos and provides a baseline for objective VQA algorithm performance.
- *Structural SIMilarity (SSIM)* is a popular method for quality assessment of still images [28], [29], that was extended to video in [30]. The SSIM index was applied frame-by-frame on the luminance component of the video [30] and the overall SSIM index for the video was computed as the average of the frame level quality scores. Matlab and Labview implementations of SSIM are available from [31].
- *Multiscale SSIM (MS-SSIM)* is an extension of the SSIM paradigm, also proposed for still images [32], that has been shown to outperform the SSIM index and many other still image quality assessment algorithms [33]. We extended the MS-SSIM index to video by applying it frame-by-frame on the luminance component of the video and the overall MS-SSIM index for the video was computed as the average of the frame level quality scores. A Matlab implementation of MS-SSIM is available for download from [31].
- *Speed SSIM* is the name we give to the VQA model proposed in [34], that uses the SSIM index in conjunction with statistical models of visual speed perception described in [35]. Using models of visual speed perception was shown to improve the performance of both PSNR and SSIM in [34]. We evaluated the performance of this framework with the SSIM index, which was shown to perform better than using the same framework with PSNR [34]. A software implementation of this index was obtained from the authors.
- *Visual Signal-to-Noise Ratio (VSNR)* is a quality assessment algorithm proposed for still images [36] and is available for download from [37]. We applied VSNR frame-by-frame on the luminance component of the video and the overall VSNR index for the video was computed as the average of the frame level VSNR scores.
- *Video Quality Metric (VQM)* is a VQA algorithm developed at the National Telecommunications and Information Administration (NTIA) [38]. Due to its excellent performance in the VQEG Phase 2 validation tests, the VQM methods were adopted by the American National Standards Institute (ANSI) as a national standard, and as International Telecommunications Union Recommendations (ITU-T J.144 and ITU-R BT.1683, both adopted in 2004). VQM is freely available for download from [39].
- *V-VIF* is the name we give to the VQA model proposed in [40] that extends the Visual Information Fidelity (VIF) criterion for still images proposed in [41] to video using temporal derivatives. A software implementation of this index was obtained from the authors.
- *MOtion-based Video Integrity Evaluation (MOVIE) index* is a VQA index that was recently developed at LIVE [42], [43]. A software implementation of MOVIE is freely available for research purposes [31]. Three different versions of the MOVIE index—the Spatial MOVIE index, the Temporal MOVIE index and the MOVIE index—were tested in our study.

A. Performance of Objective Models

We tested the performance of all objective models using two metrics—the Spearman Rank Order Correlation Coefficient (SROCC) which measures the monotonicity of the objective model prediction with respect to human scores and the Pearson Linear Correlation Coefficient (LCC) after nonlinear regression, which measures the prediction accuracy. The LCC is computed after performing a nonlinear regression on the objective VQA algorithm scores using a logistic function. We used the logistic function and the procedure outlined in [3] to fit the objective model scores to the DMOS scores.

Let Q_j represent the quality that a VQA algorithm predicts for video j in the LIVE Video Quality Database. A four-parameter, monotonic logistic function was used to fit the VQA algorithm prediction to the subjective quality scores

$$Q'_j = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-(Q_j - \beta_3/|\beta_4|)}}. \tag{7}$$

Nonlinear least squares optimization is performed using the Matlab function “nlinfit” to find the optimal parameters β that minimize the least squares error between the vector of subjective scores ($\text{DMOS}_j, j = 1, 2, \ldots 150$) and the vector of fitted objective scores ($Q'_j, j = 1, 2, \ldots, 150$). Initial estimates of the parameters were chosen based on the recommendation in [3]. We linearly rescaled VQA algorithm scores before performing the optimization to facilitate numerical convergence. The SROCC and the LCC are computed between the fitted objective scores (Q'_j) and the subjective scores (DMOS_j).

Table I (a) and (b) shows the performance of all models in terms of the SROCC and the LCC respectively for each distortion type and for the entire LIVE Video Quality Database. Scatter plots of objective scores versus DMOS for all the algorithms on the entire LIVE Video Quality Database, along with the best fitting logistic functions, are shown in Fig. 5. Our results clearly demonstrate that a carefully constructed database of videos can expose the significant limitations of PSNR as a VQA measure. All the VQA algorithms tested in our study improve upon PSNR. Speed SSIM improves upon using just the SSIM index. The best performing VQA algorithm amongst the ones tested in our study, in terms of both the SROCC and LCC after nonlinear regression, is the temporal MOVIE index. One of the three versions of the MOVIE index (Spatial MOVIE,

TABLE I

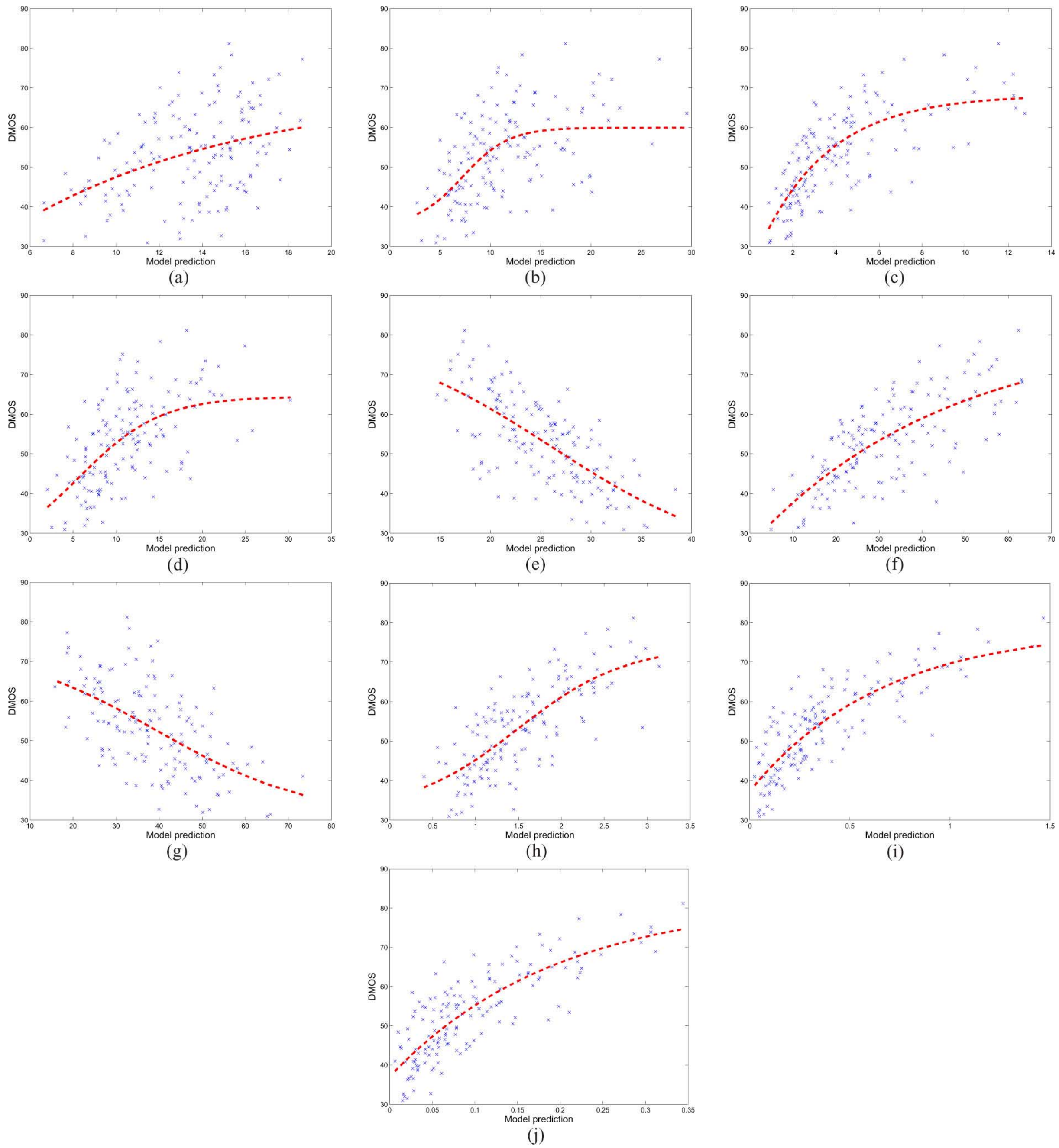
COMPARISON OF THE PERFORMANCE OF VQA ALGORITHMS. THE BEST PERFORMING ALGORITHM IS HIGHLIGHTED IN BOLD FONT FOR EACH CATEGORY. (A) SPEARMAN RANK ORDER CORRELATION COEFFICIENT, (B) LINEAR CORRELATION COEFFICIENT

Algorithm	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.4334	0.3206	0.4296	0.3588	0.3684
SSIM	0.5233	0.4550	0.6514	0.5545	0.5257
MS-SSIM	0.7285	0.6534	0.7051	0.6617	0.7361
Speed SSIM	0.5630	0.4727	0.7086	0.6185	0.5849
VSNR	0.7019	0.6894	0.6460	0.5915	0.6755
VQM	0.7214	0.6383	0.6520	0.7810	0.7026
V-VIF	0.5507	0.4736	0.6807	0.6116	0.5710
Spatial MOVIE	0.7927	0.7046	0.7066	0.6911	0.7270
Temporal MOVIE	**0.8114**	**0.7192**	**0.7797**	**0.8170**	**0.8055**
MOVIE	0.8109	0.7157	0.7664	0.7733	0.7890

(a)

Algorithm	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.4675	0.4108	0.4385	0.3856	0.4035
SSIM	0.5401	0.5119	0.6656	0.5491	0.5444
MS-SSIM	0.7170	0.7219	0.6919	0.6604	0.7441
Speed SSIM	0.5867	0.5587	0.7206	0.6270	0.5962
VSNR	0.6992	0.7341	0.6216	0.5980	0.6896
VQM	0.7325	0.6480	0.6459	0.7860	0.7236
V-VIF	0.5488	0.5102	0.6911	0.6145	0.5756
Spatial MOVIE	0.7883	0.7378	0.7252	0.6587	0.7451
Temporal MOVIE	0.8371	0.7383	**0.7920**	**0.8252**	**0.8217**
MOVIE	**0.8386**	**0.7622**	0.7902	0.7595	0.8116

(b)



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 5. Scatter plots of objective VQA scores versus DMOS for all videos in the LIVE Video Quality Database. Also shown is the best fitting logistic function. (a) PSNR. (b) SSIM. (c) MS-SSIM. (d) Speed SSIM. (e) VSNR. (f) VQM. (g) V-VIF. (h) Spatial MOVIE. (i) Temporal MOVIE. (j) MOVIE.

TABLE II

Best Performing VQA Algorithm Is Highlighted in Bold Font for Each Category. (a) Variance of the Residuals Between Individual Subjective Scores and VQA Algorithm Prediction. F-Ratios for Each Objective Model Can Be Computed as the Ratio of the Variance of the Model Residual to That of the Null Residual. F-Ratios Larger Than the Threshold F-Ratio Indicate That the Objective Model Is not Statistically Equivalent to the Null or Optimal Model. (b) Variance of the Residuals between VQA Algorithm Predictions and DMOS Values. F-Ratios to Compare Two Objective Models Can Be Computed as the Ratio of the Variances of the Model Residuals From the Two Models, With the Larger Variance Placed in the Numerator. F-Ratios Larger Than the Threshold F-Ratio Indicate That the Performance of the Objective Model in the Numerator Is Statistically Inferior to That in the Denominator

Prediction Model	Wireless	IP	H.264	MPEG-2	All Data
Null Model	105	98.61	97.73	99.24	100.18
PSNR	189.77	171.83	193.18	179.04	201.07
SSIM	180.59	164.33	166.02	165.83	184.99
MS-SSIM	156.77	140.78	159.37	152.21	153.97
Speed SSIM	174.91	159.07	157.00	157.94	177.87
VSNR	160.13	139.53	170.49	159.74	163.40
VQM	155.34	149.62	166.57	134.11	157.59
V-VIF	179.48	164.43	161.84	158.35	180.78
Spatial MOVIE	145.40	138.94	153.89	153.07	153.80
Temporal MOVIE	137.16	142.47	**142.75**	**128.72**	**139.32**
MOVIE	**136.62**	**137.38**	143.06	137.87	141.32
Number of samples	1160	870	1160	1160	4350
Threshold F-ratio	1.1015	1.1181	1.1015	1.1015	1.0512

(a)

Prediction Model	Wireless	IP	H.264	MPEG-2	All Data
PSNR	86.87	75.66	97.84	81.78	101.55
SSIM	77.46	67.91	69.98	68.24	85.36
MS-SSIM	53.07	43.58	63.15	54.30	54.15
Speed SSIM	71.64	62.48	60.73	60.16	78.20
VSNR	56.50	42.28	74.55	61.99	63.63
VQM	51.59	52.72	70.54	35.73	57.79
V-VIF	76.32	68.02	65.70	60.57	81.12
Spatial MOVIE	41.40	41.68	57.54	55.16	53.96
Temporal MOVIE	32.99	45.32	**46.14**	**30.21**	**39.41**
MOVIE	**32.41**	**40.07**	46.45	39.59	41.41
Number of samples	40	30	40	40	150
Threshold F-ratio	1.7045	1.8608	1.7045	1.7045	1.3104

(b)

Temporal MOVIE and the MOVIE index) is the best performing algorithm using SROCC or LCC as a performance indicator for each individual distortion category also. The performance of VQM, MS-SSIM and Spatial MOVIE on the LIVE Video Quality Database is comparable. Superior performance of Temporal MOVIE and MOVIE on the LIVE Video Quality Database clearly illustrates the importance of modeling visual motion perception in VQA.

B. Statistical Evaluation

The results presented in Table I (a) and (b) shows differences in the performance of different objective VQA algorithms in terms of both performance criteria. In this section, we attempt to answer the question of whether this difference in performance is statistically significant. We test the statistical significance of the results presented in Section IV-A using two different statistical tests suggested in [6]. The same tests were also used in the statistical analysis performed on the LIVE still image quality database [33]. The first is an F-test based on individual rating scores obtained from different subjects, which tests whether the performance of any objective VQA model matches the performance of humans. This test is presented in Section IV-B1. The second test is an F-test based on the errors between the average DMOS scores and model predictions, which tests whether the performance of one objective model is statistically superior to that of a competing model. This test is presented in Section IV-B2. We discuss the assumptions on which the statistical significance tests are based in Section IV-B3. See [44] for a description of statistical significance tests and F-tests.

1) F-Test Based on Individual Quality Scores: There is inherent variability amongst subjects in the quality judgment of a given video. The performance of an objective model can be, and is expected to be, only as good as the performance of humans in evaluating the quality of a given video. The optimal or "null" model obtained from the subjective study predicts the quality of a given video as the averaged Z-score across subjects, which was defined as the DMOS. The residual differences between the null model and individual quality scores assigned by each subject to a given video cannot be predicted by any objective model. Hence, the null model has a baseline residual that corresponds to the residual between individual subjective scores from different subjects and the averaged DMOS score and is given by

$$\text{Null Residual (individual ratings)} = \{z'_{ij} - \text{DMOS}_j, \quad i = 1, 2, \ldots M \text{ and } j = 1, 2, \ldots N\}. \quad (8)$$

Similar residuals can be defined for each of the objective VQA algorithms tested in the study. The residual errors between individual subjective scores and the VQA algorithm prediction of quality are given by

$$\text{Model Residual (individual ratings)} = \{z'_{ij} - Q'_j, \quad i = 1, 2, \ldots M \text{ and } j = 1, 2, \ldots N\}. \quad (9)$$

An F-test is performed on the ratio of the variance of the model residual to the variance of the null residual at 95% significance. The null hypothesis is that the variance of the model residual is equal to the variance of the null residual. A threshold F-ratio can be determined based on the number of degrees of freedom in the numerator and denominator and the significance level of the F-test. Values of the F-ratio larger than the threshold would cause us to reject the null hypothesis and conclude that the performance of the objective model is *not statistically equivalent* to the null or optimal model.

The variance of the residuals from the null model and each of the ten objective VQA models, as well as the number of samples in each category, is shown in Table II(a). The numerator and denominator degrees of freedom in the F-test is obtained by subtracting one from the number of samples. The threshold F-ratio at 95% significance is also shown in the table. None of the VQA algorithms tested in our study were found to be statistically equivalent to the null model or the theoretically optimal model corresponding to human judgment in any of the five categories (Wireless, IP, H.264, MPEG-2, or All Data). The same conclusion was reached in the VQEG Phase 2 study [6] and the LIVE still image quality study [33], wherein none of the algorithms tested in each of these studies were found to be equivalent to the theoretically optimal model. Apparently, despite significant progress, there remains considerable opportunity to improve the performance of objective VQA algorithms!

TABLE III

RESULTS OF THE F-TEST PERFORMED ON THE RESIDUALS BETWEEN MODEL PREDICTIONS AND DMOS VALUES. EACH ENTRY IN THE TABLE IS A CODEWORD CONSISTING OF FIVE SYMBOLS. THE SYMBOLS CORRESPOND TO "WIRELESS", "IP," "H.264," "MPEG-2" AND "ALL DATA" IN THAT ORDER. A SYMBOL VALUE OF "1" INDICATES THAT THE STATISTICAL PERFORMANCE OF THE VQA MODEL IN THE ROW IS SUPERIOR TO THAT OF THE MODEL IN THE COLUMN. A SYMBOL VALUE OF "0" INDICATES THAT THE STATISTICAL PERFORMANCE OF THE MODEL IN THE ROW IS INFERIOR TO THAT OF THE MODEL IN THE COLUMN AND "-" INDICATES THAT THE STATISTICAL PERFORMANCE OF THE MODEL IN THE ROW IS EQUIVALENT TO THAT OF THE MODEL IN THE COLUMN. NOTICE THAT THE MATRIX IS SYMMETRIC AND THAT THE CODEWORDS AT TRANSPOSE LOCATIONS IN THE MATRIX ARE BINARY COMPLEMENTS OF EACH OTHER. M1 THROUGH M10 ARE PSNR, SSIM, MS-SSIM, SPEED SSIM, VSNR, VQM, V-VIF, SPATIAL MOVIE, TEMPORAL MOVIE, AND MOVIE RESPECTIVELY

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	- - - - -	- - - - -	- - - - 0	- - - - -	- - - - 0	- - - 0 0	- - - - -	0 - - - 0	0 - 0 0 0	0 - 0 0 0
M2	- - - - -	- - - - -	- - - - 0	- - - - -	- - - - 0	- - - 0 0	- - - - -	0 - - - 0	0 - - 0 0	0 - - 0 0
M3	- - - - 1	- - - - 1	- - - - -	- - - - 1	- - - - -	- - - - -	- - - - 1	- - - - -	- - - 0 0	- - - - -
M4	- - - - -	- - - - -	- - - - 0	- - - - -	- - - - -	- - - - 0	- - - - -	0 - - - 0	0 - - 0 0	0 - - - 0
M5	- - - - 1	- - - - 1	- - - - -	- - - - -	- - - - -	- - - 0 -	- - - - -	- - - - -	0 - - 0 0	0 - - - 0
M6	- - - 1 1	- - - 1 1	- - - - -	- - - - 1	- - - 1 -	- - - - -	- - - - 1	- - - - -	- - - - 0	- - - - 0
M7	- - - - -	- - - - -	- - - - 0	- - - - -	- - - - -	- - - - 0	- - - - -	0 - - - 0	0 - - 0 0	0 - - - 0
M8	1 - - - 1	1 - - - 1	- - - - -	1 - - - 1	- - - - -	- - - - -	1 - - - 1	- - - - -	- - - 0 0	- - - - -
M9	1 - 1 1 1	1 - - 1 1	- - - 1 1	1 - - 1 1	1 - - 1 1	- - - - 1	1 - - 1 1	- - - 1 1	- - - - -	- - - - -
M10	1 1 1 1 1	1 - - 1 1	- - - - -	1 - - - 1	1 - - - 1	- - - - 1	1 - - - 1	- - - - -	- - - - -	- - - - -

2) F-Test Based on Average Quality Scores: The residual error between the quality predictions of an objective VQA model and the DMOS values on the LIVE Video Quality Database can be used to test the statistical superiority of one VQA model over another. The residual errors between the objective algorithm prediction and the DMOS value is given by

$$\text{Model Residual (average ratings)} = \{Q'_j - \text{DMOS}_j, \quad j = 1, 2, \ldots N\}. \quad (10)$$

An F-test is performed on the ratio of the variance of the residual error from one objective model to that of another objective model at 95% significance level. The null hypothesis states that variances of the error residuals from the two different objective models are equal. The variance of the residual errors between model predictions and the DMOS for all the objective models tested in our study for all the categories are shown in Table II(b). The F-ratio is always formed by placing the objective model with the larger residual error variance in the numerator. Threshold F-ratios can be determined based on the number of samples in each category and the significance level. The threshold F-ratio and the number of samples in each category are also listed in Table II(b). An F-ratio ratio larger than the threshold indicates that the performance of the VQA algorithm in the numerator of the F-ratio is statistically inferior to that of the VQA algorithm in the denominator. The results of the statistical significance test are reported in Table III.

To summarize the results in Table III, the performance of Temporal MOVIE and MOVIE is statistically superior to that of PSNR, SSIM, Speed SSIM, VSNR, and V-VIF and the performance of Spatial MOVIE is superior to that of PSNR, SSIM, VSNR, and V-VIF on the wireless dataset. The performance of all algorithms are statistically equivalent on the IP dataset. The performance of Temporal MOVIE and MOVIE are statistically superior to PSNR on the H.264 dataset. The performance of VQM is superior to PSNR, SSIM, and VSNR and the performance of MOVIE is superior to PSNR and SSIM on the MPEG-2 dataset. Additionally, the performance of Temporal MOVIE is superior to PSNR, SSIM, MS-SSIM, Speed SSIM, VSNR, and V-VIF on the MPEG-2 dataset.

The performance of Temporal MOVIE, which is the best performing algorithm on the entire LIVE Video Quality Database, is statistically superior to the performance of all algorithms tested in the study, with the exception of MOVIE. The MOVIE index is statistically superior to PSNR, SSIM, Speed SSIM, VSNR and V-VIF on the entire LIVE Video Quality Database. Spatial MOVIE, MS-SSIM , and VQM are superior to PSNR, SSIM, Speed SSIM, and V-VIF on the entire LIVE Video Quality Database. Finally, the performance of VQM is superior to that of PSNR and SSIM on the entire LIVE Video Quality Database.

Assumptions of the F-Test: The F-test that we use assumes that the residuals are independent samples from a normal distribution and is fairly robust to this assumption [44]. For additional verification of the robustness of the F-tests to the underlying assumptions, we also performed bootstrapped F-tests on both the individual quality scores and the average quality scores [45]. For instance, bootstrapped F-tests on the average quality scores were performed by selecting N values from the vectors of model residuals in (10) randomly with resampling for each of the two models under test and computing the F-ratio. This procedure is repeated 10 000 times to obtain the sampling distribution of the F-ratio. We visually verified that the sampling distribution of the F-ratio is shifted to the right of 1 for all cases where statistical significance was established. Due to space limitations, we only show the sampling distribution of the F-ratio on the entire LIVE Video Quality Database for each of the six models whose performance is statistically superior to PSNR in Fig. 6.

For additional verification of the assumptions of the F-test, we performed another simulation where we generated N independent samples from a standard normal distribution with the same mean and variance as the vector of model residuals in (10). The F-ratio was then computed between each pair of objective models. This procedure was repeated 10 000 times for each pair to obtain the sampling distribution of the F-ratio when the assumptions of the F-test are exactly met. The resulting sampling distribution is also shown in Fig. 6 in dotted lines. It is seen that the two sampling distributions are quite close to each other, which shows that any deviations of the distribution of the residual data from the assumption of independent and Gaussian residuals do not affect the results of the statistical tests greatly.

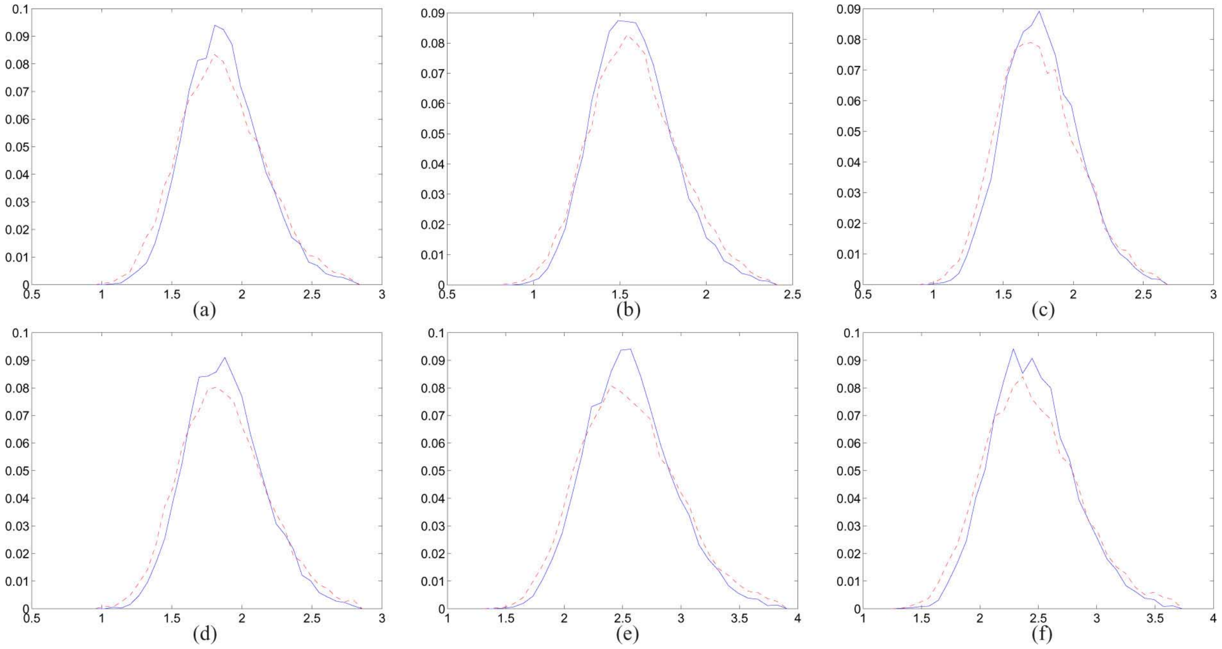


Fig. 6. Sampling distribution of the F-ratio obtained using bootstrap simulations for the F-test based on average quality scores. Sampling distributions are shown for all VQA models that are statistically superior to PSNR on the entire LIVE Video Quality Database. Note that the sampling distributions are shifted to the right of 1. Also shown in dotted lines is the sampling distribution of the F-ratio when random samples are generated to exactly satisfy the assumptions of the F-test. Note that the sampling distribution obtained from the data and the sampling distribution obtained from simulated data that satisfy the F-test assumptions are very similar. (a) PSNR versus MS-SSIM. (b) PSNR versus VSNR. (c) PSNR versus VQM. (d) PSNR versus Spatial MOVIE. (e) PSNR versus Temporal MOVIE. (f) PSNR versus MOVIE.

This simulation was also performed for the F-tests based on individual quality scores with identical conclusions.

C. Discussion of Results

The intention of this study has been to provide an independent, academic VQA resource that is freely available to download, free from commercial interests, broadly representative of applications, and that will be continuously vital, since the database will be updated over time. Future human studies are also planned that will extend the scope of the current study.

The study has been a rather large undertaking. Of course, the results of the human study and of the algorithm comparisons do not represent a final statement, since in coming years new theories and algorithms will continue to be developed in this exciting area, existing algorithms will be improved, and some unavailable (proprietary) algorithms may be offered for comparison (we continue our efforts to obtain these). As video applications continue to evolve, the set of distortions to be considered as "representative" will naturally change over time as well. New developments will be posted on the LIVE VQA website [8] on a regular basis.

The results that we obtained here affirm long-held beliefs regarding the failure of "classical" measures of video "quality" to predict the human sense of quality. Most notably, the long-used PSNR has been shown to perform very poorly against human subjectivity, far worse than any of the perceptually relevant algorithms considered. We hope that this result helps lay to rest, at long last, the notion that the PSNR is a reliable predictor, measure, or optimizer of video (or image) quality—at least for applications where humans are the video "receivers." If we succeed in hastening the demise of the PSNR, then it will, perhaps, be the most gratifying and important product of this effort.

The correlation study comparing the various VQA algorithms against the large set of human data produced a number of useful results and some surprising ones as well. Good performance of two of the algorithms (MS-SSIM [32] and the VQM from NTIA [38]) affirm both of these algorithms as extremely practical and well-suited to benchmark video processing algorithms, especially since both algorithms do not perform computationally intensive operations such as motion estimation. Since both algorithms are freely available for download (although VQM is restricted for commercial use) [27], [39], these can be easily used to analyze the performance of a video processing algorithm, provided that the performance simulations have available a reference for comparison.

The notion that using computed motion information can improve VQA algorithm performance is strongly validated by the study. For example, "Speed SSIM" [34] exhibits substantially improved performance relative to simple (single-scale) SSIM [29]. One wonders at how well "Speed SSIM" might perform if made multiscale, which would require some nontrivial design. Nevertheless, the distinction in performance between simple SSIM and MS-SSIM suggests that this might be a fruitful development. Likewise, the still image algorithm VSNR [36] also performed well, suggesting that a future version of this algorithm that seeks to incorporate temporal information should be encouraged.

The Temporal MOVIE index, described in detail in [42] and [43], yielded the best overall performance and is statistically superior to all other algorithms tested in this study with the exception of MOVIE. Before discussing this performance, we note that the MOVIE algorithm tested on this database is unchanged from the one reported in the literature and successfully tested on the VQEG database. The algorithm was "frozen" before the data from the human studies provided here were completely captured, analyzed, and used to compare algorithms. As described in [42] and [43], the few parameters (three masking constants) in the MOVIE index were selected to take values equal to the nearest order of magnitude of an appropriate energy term. While it is possible that parameter "fiddling" could improve any VQA algorithm (for example, the VQM algorithm has been trained on the VQEG FRTV Phase 1 database as part of the process of selecting its many parameters), this has not been done with the MOVIE index.

Instead, the success of the MOVIE index lies in two directions: first, the use of perceptually relevant models of human visual perception in space and time. As described in [43], MOVIE utilizes specific (Gabor receptive field) models of cortical area V1 to dissemble video data into multiscale space-time primitives. The Gabor receptive field model has produced dominant approaches to many fundamental vision engineering problems, such as texture analysis [46], [47], motion analysis [48], computational stereo [49], and human biometrics [50], [51]. MOVIE also uses a specific model of the relatively well-understood extracortical area V5 (also known as area MT) to effect a biologically plausible model of visual motion processing [52]. Using these models, MOVIE deploys SSIM-like multiscale processing to compute local scale-space comparisons that can be supported from an information-theoretic viewpoint under natural scene statistical models [53].

Looking at the breakdown of MOVIE into its spatial and temporal components, it may be observed that Spatial MOVIE attains a level of performance very similar to that of MS-SSIM and VQM—overall, in nearly every category and statistically. Indeed, Spatial MOVIE may be viewed as a perceptual matched version of MS-SSIM, owing to its use of spatio-temporal basis functions. Temporal MOVIE performs considerably better than Spatial MOVIE and every other algorithm tested in our study and the improvement is shown to be statistically significant, despite not being tuned to detect spatial distortions (of which the database contains many). MOVIE also shows excellent performance and is statistically superior to PSNR, SSIM, Speed SSIM, VSNR, VQM, and V-VIF. We believe that these results powerfully illustrate the need for modeling visual motion processing in VQA. It is interesting that the performance of Temporal MOVIE is better than that of MOVIE overall. However, this difference in performance is not statistically significant and further, MOVIE performs better than Temporal MOVIE on the wireless and IP videos in terms of LCC and on the VQEG database [43].

In our view, it is plausible that MOVIE might approach the limits of performance that might be obtained by VQA algorithms without taking into account other factors, such as human attention, foveation, and salience [54]. These are topics for future studies.

Broadly, this study shows that there are a number of algorithms that perform significantly better than traditional methods with a high degree of statistical confidence. We have the opinion that these and future algorithms should play an increasingly important role in the benchmarking and design of video processing systems.

V. Conclusions and Future Work

A subjective study to evaluate the effects of present generation video compression and communication technologies on the perceptual quality of digital video was presented. This study included 150 videos derived from ten reference videos using four distortion types and were evaluated by 38 subjects. The resulting LIVE Video Quality Database is unique in terms of content and distortion and is publicly available for research purposes [8]. We presented an evaluation of the performance of several publicly available objective VQA models on this database.

A distinguishing feature of our database was that the distortion strengths were adjusted perceptually to test the ability of VQA models to perform consistently well across content types. Two of the distortion types in our database resulting from video transmission through lossy wireless and IP networks cause distortions that are transient, both spatially and temporally. This is another distinguishing and important aspect of the database. VQA algorithms need to be able to account for such transient distortions. Regarding the evaluation of objective quality indexes using the linear (Pearson) correlation (LCC), a logistic function was used to fit the data to account for nonlinearities in the objective model. It can be argued that it would be convenient for an objective model to have a linear relationship with subjective quality judgments, since it would allow for easier interpretation and use of the VQA algorithm. Of course, VQA and still image quality assessment algorithms generally do not exhibit linear behavior relative to human subjective judgments (Section IV-A). Nonlinearities in the objective model can be accounted for by calibration within the model, with the added caveat that subjective judgment of quality can vary with subjective data processing, context, calibration, and range of subjective qualities being considered. While linearity of a model relative to subjectivity is convenient for interpretation, in our view, correlation measures that do not rely on any linearity assumptions, such as SROCC, that are independent of any function mapping between the objective and subjective scores are particularly useful for direct algorithm comparisons.

As part of our study, we also recorded quality scores in continuous time provided by the subject as they are viewing the video. This provides a description of the quality of the video as a function of time. We intend to make use of this data in the future to design pooling strategies for objective VQA algorithms that can correlate with human data scores. The single stimulus subjective testing paradigm with hidden reference removal used in the LIVE Video Quality Database makes it amenable to testing the performance of no-reference VQA algorithms. No-reference VQA is a far less mature field than full-reference VQA and the focus to date has largely been on application-specific metrics that measure the perceptual strength of specific distortions typical in applications such as compression, network transmission of video and so on. We intend to work on the elusive goal of generic no-reference VQA in the future and hope that the LIVE Video Quality Database will prove valuable in advancing the state of the art in this field also.

Acknowledgment

The authors would like to thank J. Perry of the Center for Perceptual Systems (CPS) at the University of Texas at Austin for providing the XGL toolbox and for several helpful discussions on developing the subjective study interface.

References

[1] LIVE Image Quality Assessment Database, 2003 [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm

[2] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "A subjective study to evaluate video quality assessment algorithms," in *Proc. SPIE—Human Vision and Electronic Imaging*, 2010.

[3] 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

[4] A. Webster, "Progress and future plans for VQEG," in *ETSI STQ Workshop on Multimedia Quality of Service*, 2008 [Online]. Available: http://portal.etsi.org/docbox/Workshop/2008/2008_06_STQWORKSHOP/VQEG_ArthurWebster.pdf

[5] Video Quality Experts Group, 2003 [Online]. Available: http://www.its.bldrdoc.gov/vqeg/

[6] Final VQEG Report on the Validation of Objective Models of Video Quality Assessment The Video Quality Experts Group, 2003 [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII

[7] A. Webster, Final Report From the Video Quality Experts Group on the Validation of Objective Models of Multimedia Quality Assessment, Phase 1 2008 [Online]. Available: ftp://vqeg.its.bldrdoc.gov/Documents/Projects/multimedia/MM_Final_Report/

[8] LIVE Video Quality Database, 2009 [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html

[9] Tech. Univ. Munich [Online]. Available: ftp://ftp.ldv.e-technik.tu-muenchen.de/pub/test_sequences/

[10] *Generic coding of moving pictures and associated audio information—Part 2: Video*, ITU-T Recommendation H.262—ISO/IEC 13818-2 (MPEG-2), IUT and ISO/IEC JTC1, Nov. 2004 [Online]. Available: http://standards.iso.org/ittf/PubliclyAvailableStandards/c039486ISOIEC13818-52005ReferenceSoftware.zip

[11] H.264/AVC Software Coordination, 2007 [Online]. Available: http://iphome.hhi.de/suehring/tml/

[12] S. Wenger, "H.264/AVC over IP," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 645–656, Jul. 2003.

[13] Proposed Error Patterns for Internet Experiments, 1999 [Online]. Available: http://ftp3.itu.ch/av-arch/video-site/9910_Red/q15i16.zip

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

[15] T. Stockhammer, M. 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.

[16] Common Test Conditions for RTP/IP Over 3GPP/3GPP2, 1999 [Online]. Available: http://ftp3.itu.ch/av-arch/video-site/0109_San/VCEG-N80_software.zip

[17] RRNR-TV Group Test Plan, 2008 [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/rrnr-tv/

[18] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Proc. SPIE—Visual Communications and Image Processing*, 2003.

[19] The XGL Toolbox, 2008 [Online]. Available: http://128.83.207.86/~jsp/software/xgltoolbox-1.0.5.zip

[20] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," in *Proc. SPIE—Advanced Image and Video Communications and Storage Technologies*, 1995.

[21] Int. Telecommun. Union, Methodology for the Subjective Assessment of the Quality of Television Pictures ITU-R Recommendation BT.500-11, Tech. Rep., 2000.

[22] Tektronix [Online]. Available: http://www.tek.com/products/video_test/pqa500/

[23] Opticom [Online]. Available: http://www.opticom.de/technology/pevq_video-quality-testing.html

[24] Symmetricom [Online]. Available: http://qoe.symmetricom.com/

[25] SwissQual [Online]. Available: http://www.swissqual.com/Algorithms.aspx

[26] Kwill Corp. [Online]. Available: http://www.kwillcorporation.com/products/VP21H.html

[27] Video quality experts group [Online]. Available: http://www.its.bldrdoc.gov/vqeg/links/links.php

[28] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[29] 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.

[30] 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.

[31] LIVE Software Releases [Online]. Available: http://live.ece.utexas.edu/research/Quality/index.htm

[32] Z. Wang, E. P. Simoncelli, A. C. Bovik, and M. Matthews, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals, Systems and Computers*, 2003, pp. 1398–1402.

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

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

[35] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nat. Neurosci.*, vol. 9, no. 4, pp. 578–585, Apr. 2006.

[36] 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.

[37] MeTriX MuX Visual Quality Assessment Package [Online]. Available: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[38] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[39] VQM [Online]. Available: http://www.its.bldrdoc.gov/n3/video/VQM_software.php

[40] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *1st Int. Conf. Video Processing and Quality Metrics for Consumer Electronics*, 2005.

[41] 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.

[42] K. Seshadrinathan and A. C. Bovik, "Motion-based perceptual quality assessment of video," in *Proc. SPIE—Human Vision and Electronic Imaging*, 2009.

[43] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[44] D. C. Howell, *Statistical Methods for Psychology*. Belmont, CA: Wadsworth, 2007.

[45] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap* , ser. Monographs on Statistics and Applied Probability. London, U.K.: Chapman and Hall, 1993.

[46] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, Jan. 1990.

[47] A. C. Bovik, N. Gopal, T. Emmoth, and A. Restrepo, "Localized measurement of emergent image frequencies by Gabor wavelets," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 691–712, Feb. 1992.

[48] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vision*, vol. 5, no. 1, pp. 77–104, 1990.

[49] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin, "Phase-based disparity measurement," *Comput. Vis., Graphics, Image Process.: Image Underst.*, vol. 53, no. 2, pp. 198–210, 1991.

[50] J. Daugman, "How iris recognition works," in *The Handbook of Image and Video Processing*, A. C. Bovik, Ed. New York: Elsevier, 2005.

[51] L. Wiskott, J. M. Fellous, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, Jul. 1997.

[52] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vision Res.*, vol. 38, no. 5, pp. 743–761, Mar. 1998.

[53] K. Seshadrinathan and A. C. Bovik, "Unifying analysis of full reference image quality assessment.," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 1200–1203.

[54] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Trans. Image Process*, vol. 17, no. 4, pp. 564–573, 2008.

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

She is currently a System Engineer with Intel Corporation, Chandler, AZ. Her 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. She was Assistant Director of the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin from 2005–2008.

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 at Austin.

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

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

Alan Conrad Bovik (S'80–M'81–SM'89–F'96) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, in 1980, 1982, and 1984, respectively.

He is currently the Curry/Cullen Trust Endowed Professor at the University of Texas at Austin, where he is the Director of the Laboratory for Image and Video Engineering (LIVE) in the Center for Perceptual Systems. His research interests include image and video processing, computational vision, digital microscopy, and modeling of biological visual perception. He has published over 450 technical articles in these areas and holds two U.S. patents. He is also the author of *The Handbook of Image and Video Processing* (Elsevier, 2005, 2nd ed.) and *Modern Image Quality Assessment* (Morgan & Claypool, 2006).

Dr. Bovik has received a number of major awards from the IEEE Signal Processing Society, including: the Education Award (2007); the Technical Achievement Award (2005); the Distinguished Lecturer Award (2000); and the Meritorious Service Award (1998). He is also a recipient of the Distinguished Alumni Award from the University of Illinois at Urbana-Champaign (2008), the IEEE Third Millennium Medal (2000), and two journal paper awards from the International Pattern Recognition Society (1988 and 1993). He is a Fellow of the Optical Society of America and the Society of Photo-Optical and Instrumentation Engineers. He has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996–1998; Editor-in-Chief IEEE TRANSACTIONS ON IMAGE PROCESSING, 1996–2002; Editorial Board PROCEEDINGS OF THE IEEE, 1998–2004; Series Editor for *Image, Video, and Multimedia Processing* (Morgan and Claypool, 2003–present); and Founding General Chairman, First IEEE International Conference on Image Processing, Austin, TX, November 1994. He is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial, and academic institutions.

Lawrence K. Cormack is from Socorro, NM. He received the B.S. degree with Highest Honors in psychology from the University of Florida, Gainesville, in 1986, and the Ph.D. degree in physiological optics (now vision science) from the University of California, Berkeley in 1992.

He joined the Psychology Department at the University of Texas at Austin (UTA) in the fall of 1992 and has been there ever since. At UTA, he is also an active member of the Center for Perceptual Systems, the Institute for Neuroscience, and the Imaging Research Center. His research interests are primarily in the brain mechanisms underlying vision, including the basic questions of: 1) the way depth and motion are perceived (i.e., dynamic 3-D perception) and 2) the role that the structure of the natural environment has played in the evolution of the visual system.