A Subjective Study to Evaluate Video Quality Assessment Algorithms

Kalpana Seshadrinathan[a], Rajiv Soundararajan[b], Alan C. Bovik[b] and Lawrence K. Cormack[b]

[a] Intel Corporation, Chandler, AZ - USA.

[b] The University of Texas at Austin, Austin, TX - USA.

ABSTRACT

Automatic methods to evaluate the perceptual quality of a digital video sequence have widespread applications wherever the end-user is a human. Several objective video quality assessment (VQA) algorithms exist, whose performance is typically evaluated using the results of a subjective study performed by the video quality experts group (VQEG) in 2000. There is a great need for a free, publicly available subjective study of video quality that embodies state-of-the-art in video processing technology and that is effective in challenging and benchmarking objective VQA algorithms. In this paper, we present a study and a resulting database, known as the LIVE Video Quality Database, where 150 distorted video sequences obtained from 10 different source video content were subjectively evaluated by 38 human observers. Our study includes videos that have been compressed by MPEG-2 and H.264, as well as videos obtained by simulated transmission of H.264 compressed streams through error prone IP and wireless networks. The subjective evaluation was performed using a single stimulus paradigm with hidden reference removal, where the observers were asked to provide their opinion of video quality on a continuous scale. We also present the performance of several freely available objective, full reference (FR) VQA algorithms on the LIVE Video Quality Database. The recent MOtion-based Video Integrity Evaluation (MOVIE) index emerges as the leading objective VQA algorithm in our study, while the performance of the Video Quality Metric (VQM) and the Multi-Scale Structural SIMilarity (MS-SSIM) index is noteworthy. The LIVE Video Quality Database is freely available for download[1] and we hope that our study provides researchers with a valuable tool to benchmark and improve the performance of objective VQA algorithms.

Keywords: video quality, quality assessment, subjective study, LIVE Video Quality Database, full reference, MOVIE

1. INTRODUCTION

Digital video applications such as digital television, digital cinema, video teleconferencing, IPTV, mobile TV, streaming videos over the Internet and wireless networks target a human observer as the end user of the video. We use "VQA" to refer solely to the perceptual quality of a video (where the end user is a human observer) and the objective is to evaluate the quality of a video as perceived by an average human observer. VQA methods play a critical role in the design of video communication systems, from the point of video acquisition until the video is displayed to the human.

Subjective VQA deals with methods that utilize human subjects to perform the task of assessing visual quality. Since we are interested in human opinions of quality, subjective VQA is the only reliable method of performing VQA. However, it is impossible to subjectively assess the quality of each and every video that is come across in an application. Due to inherent variability in quality judgment amongst human observers, multiple subjects who are representative of the target audience are required to participate in a subjective study. Video quality is affected by viewing conditions such as ambient illumination, display device, viewing distance and so on and subjective studies have to be conducted in a carefully controlled environment. Subjective VQA is cumbersome and expensive, but is valuable in providing ground truth data for the evaluation of automatic or objective VQA algorithms. Objective VQA algorithms eliminate human involvement and automatically predict the visual quality of an input video. We focus on full reference (FR) VQA algorithms in this paper that assume the availability of a pristine, original, reference video in addition to the test video whose quality is to be evaluated.

Currently, the performance of objective VQA algorithms is largely evaluated on the publicly available Video Quality Experts Group (VQEG) FRTV Phase-I database.[2] The VQEG database was published in 2000 and the

distortions in the test videos (for instance, MPEG-2 and H.263 compression) are not representative of present generation video encoders (H.264) and communication systems. Videos in the VQEG study are interlaced and interlacing causes artifacts in the reference videos, requires modifications to objective VQA systems to handle interlacing, and is not representative of multimedia and other applications that use progressive videos. 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 $>$ 3Mbps). Overall, the VQEG videos exhibit poor perceptual separation, making it difficult to distinguish the performance of VQA algorithms. These limitations of the VQEG database greatly motivate our work. In this paper, we first present a subjective study that included 10 raw naturalistic reference videos and 150 distorted videos obtained from the references using four real world distortion types. The quality of each video was evaluated by 38 subjects using a single stimulus paradigm 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 popular and widely used LIVE Image Quality Database for still images.[1,3] We then present an evaluation of the performance of leading, publicly available objective VQA algorithms on our database. The LIVE Video Quality Database is freely available for download and we hope that it provides researchers with a valuable tool to benchmark and improve the performance of objective VQA algorithms.[1]

The subjective study of video quality is described in Section 2. We discuss the performance evaluation of several publicly available FR VQA algorithms on our database in Sections 3 and 4. We conclude this paper in Section 5 with a discussion of future work.

2. SUBJECTIVE STUDY OF VIDEO QUALITY

2.1 Source Sequences

We used ten uncompressed videos of natural scenes as source videos (as opposed to animation, graphics, text etc) that are freely available for download from the Technical University of Munich.[4] The digital videos are provided in uncompressed YUV 4:2:0 format (which guarantees that the reference videos are distortion free) and do not contain audio. We only used progressively scanned videos to avoid problems associated with video de-interlacing. Although the videos in the Munich database were captured in High Definition (HD) format, we downsampled all videos to a resolution of 768X432 pixels due to resource limitations in displaying the videos. We chose this resolution to ensure that the aspect ratio of the HD videos are maintained, thus minimizing visual distortions. We downsampled each video frame by frame using the "imresize" function in Matlab using bicubic interpolation to minimize distortions due to aliasing. Nine out of ten videos were 10 seconds long, while the 10^{th} video was 8.68 seconds long. Seven sequences had a frame rate of 25 frames per second, while the remaining three had a frame rate of 50 frames per second. The videos were diverse in content and included a wide range of objects, textures, motions and camera movements.

2.2 Test Sequences

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. We created a total of 15 test sequences from each of the reference sequences using four different distortion processes - MPEG-2 compression (4 test videos per reference), H.264 compression (4 test videos per reference), lossy transmission of H.264 compressed bitstreams through simulated IP networks (3 test videos per reference) and lossy transmission of H.264 compressed bitstreams through simulated wireless networks (4 test videos per reference).

The MPEG-2 compressed videos were created by compressing the reference to different bit rates using the MPEG-2 reference software.[5] H.264 compressed videos were created using the JM reference software (Version 12.3) made available by the Joint Video Team (JVT).[6] Video communication of H.264 videos over IP networks has been studied[7] and our design of a video communication system that simulate losses in H.264 video streams transmitted over IP networks was based on this study. The video sequences subjected to errors in the IP environment contained between one and four slices per frame; we used these two options since they result in packet sizes that are typical in IP networks. Four IP error patterns from real-world experiments supplied by the

Video Coding Experts Group (VCEG), with loss rates of 3%, 5%, 10% and 20%, were used.[8] We created test videos by dropping packets specified in the error pattern from an H.264 compressed packetized video stream. The resulting H.264 bitstream was then decoded using the JM reference software[6] and the losses were concealed using the built-in error concealment mechanism (mode 2 - motion copy). Video communication of H.264 videos over wireless networks has also been studied[9] and our design of a video communication system to simulate losses in H.264 streams transmitted over wireless networks was based on this study. A packet from an H.264 compressed and packetized bitstream transmitted over a wireless channel is susceptible to bit errors due to attenuation, shadowing, fading and multi-user interference in wireless channels. We assume that a packet is lost even if it contains a single bit error, an assumption that is often made in practice.[9] We simulated errors in wireless environments using bit error patterns and software available from the VCEG.[10] Decoding and error concealment for the wireless simulations were identical to the IP simulations.

Compression systems such as MPEG-2 and H.264 produce fairly uniform distortions in the video, both spatially and temporally. Network losses, however, cause *transient* distortions in the video. MPEG-2 and H.264 compressed videos exhibit compression artifacts such as blocking, blur, ringing and motion compensation mismatches around object edges. Videos obtained from lossy transmission through IP and wireless networks exhibit errors that are restricted to small regions of the video that correspond to the lost packets. The error concealment mechanisms cause distinct visual artifacts in the wireless and IP videos, that are very different from compression related distortions. Figure 1 shows a frame from videos obtained from each of the four distortion categories in our database.

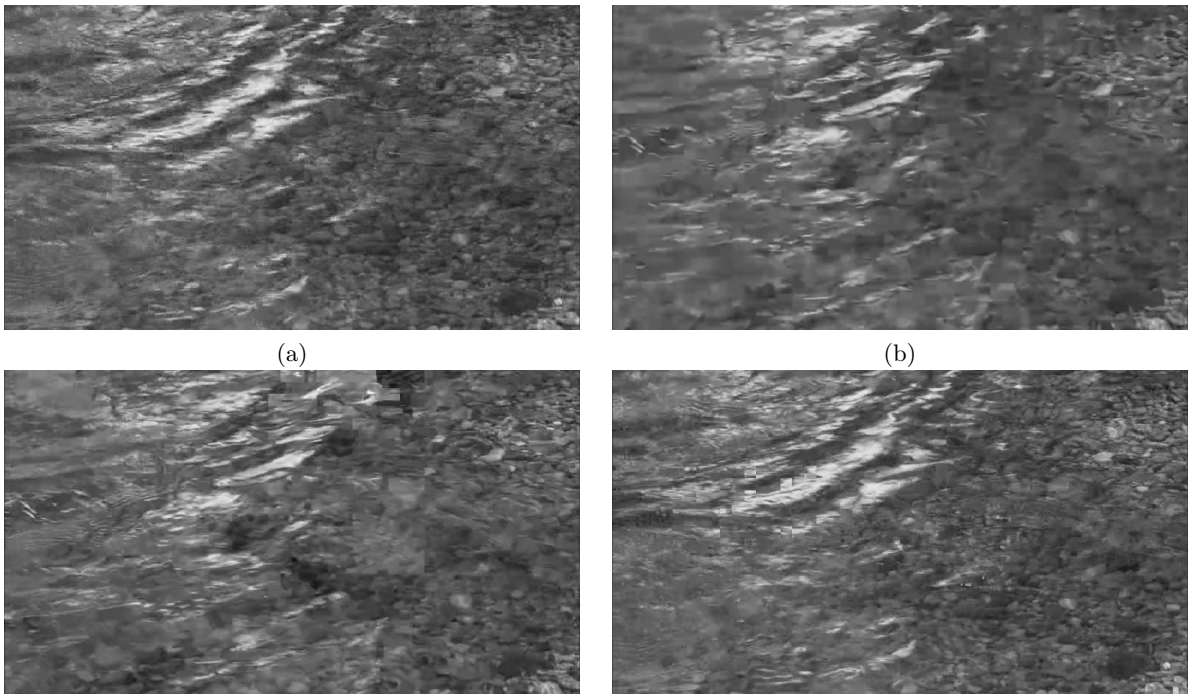
The distortion strengths were adjusted manually so that the videos obtained from each source and each distortion category spanned a set of contours of equal visual quality. A large set of videos were generated and viewed by the authors and a subset of these videos that spanned the desired visual quality 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"). Four MPEG-2 compressed versions of each reference video 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 the reference video. The same selection procedure is then repeated for every reference video. Note that all "Excellent" videos are designed to have the approximate same visual quality, across reference videos and across the four distortion types. Thus, 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 highly effective in challenging and distinguishing the performance of objective VQA algorithms as shown in Section 3.

2.3 Subjective Testing

We adopted a single stimulus paradigm with hidden reference removal and a continuous quality scale to obtain subjective quality ratings for the video database. The subject indicates the quality of each video he views on a continuous quality scale displayed on the screen as a slider bar after the presentation of the video. The subject is also presented with the reference video during the study which he scores (the references are interspersed in the study and the subject is not aware that he is scoring the reference). The scores assigned to the reference are then used to compute Difference Mean Opinion Scores (DMOS) between each test video and the corresponding reference.

All the videos in our study were viewed by each subject, which required one hour of the subject's time. To minimize the effects of viewer fatigue, we conducted the study in two sessions of a half hour each. We prepared playlists for each subject by arranging the 150 test videos in a random order. We ensured that the subjects did not view successive presentations of test videos that were obtained from the same reference, to avoid contextual and memory effects in their judgment of quality. Additionally, we included each reference in both sessions in the hidden reference removal process to avoid any session dependent effects from affecting the results of the study.

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.[11] The XGL toolbox allows precise presentation of psychophysical stimuli to human observers and allowed us to ensure that the timing of the video playback was precise. The videos were viewed by the subjects on a Cathode Ray Tube (CRT) monitor and



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

Figure 1: (a) MPEG-2 compressed frame (b) H.264 compressed frame (c) IP loss simulated frame (d) Wireless loss simulated frame

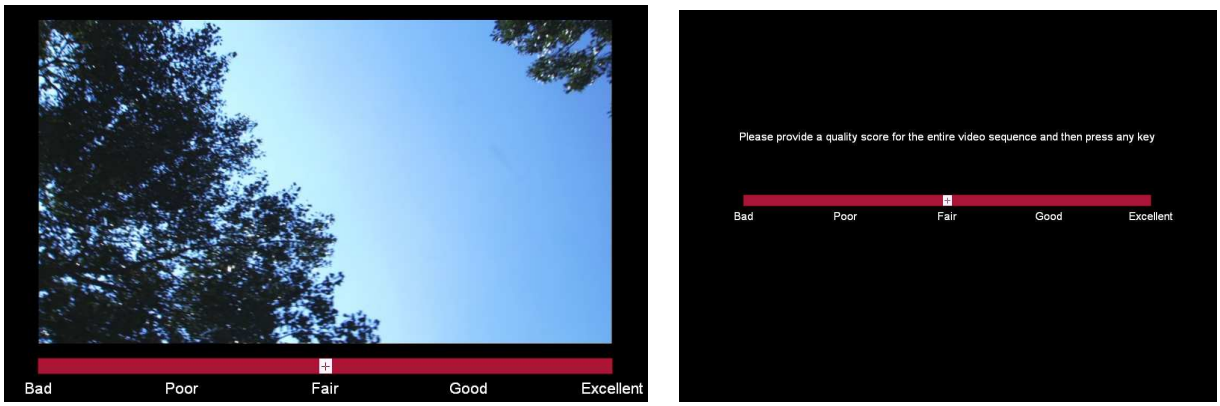
the entire study was conducted using the same monitor, which was calibrated using the Monaco Optix XR Pro device. Since the videos had a low frame rate (25 and 50 Hz), we set the monitor resolution to 100 Hz to avoid artifacts due to monitor flicker. Each frame of the 25 and 50 Hz videos were displayed for 2 and 4 monitor refresh cycles respectively. 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 subject could provide his opinion of quality by moving his mouse along the slider. The quality scale had five, equally spaced labels - "Excellent", "Good", "Fair", "Poor" and "Bad" - marked on it to help the subject. Screenshots from the subjective study interface are shown in Fig. 2.

We used 38 students at The University of Texas at Austin as subjects in our study. The subjects viewed a short training session to familiarize themselves with the subjective experiment, user interface and the range of visual quality that they could expect in the study. The training content was different from the videos in our study and were impaired using the same distortion types.

2.4 Processing of Subjective Scores

Let s_{ijk} denote the score assigned by subject i to video j in session $k = \{1, 2\}$. 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.

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



(a) (b)

Figure 2: (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 difference scores for the reference videos are 0 in both sessions and are removed. The difference scores are then converted to Z-scores per session:[12]

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

$$z_{ijk} = \frac{d_{ijk} - \mu_{ik}}{\sigma_{ik}} \tag{3}$$

where N_{ik} is the number of test videos seen by subject i in session k.

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 $\{z_{ij}\}$. Scores from unreliable subjects are discarded using the procedure specified in the ITU-R BT 500.11 recommendation.[13] 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.[13] 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.

The Z-scores were then linearly rescaled to lie in the range $[0, 100]$. Finally, the DMOS of each video is computed as the mean of the rescaled Z-scores from the 29 remaining subjects after subject rejection.

3. PERFORMANCE OF OBJECTIVE VQA ALGORITHMS

The performance of several publicly available objective VQA models was evaluated on our database. Many popular VQA algorithms are licensed and sold for profit and are not 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,[14,15] that was extended to video.[16] The SSIM index was applied frame-by-frame on the luminance component of the video[16] 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 freely available for download.[17]
- *Multi-scale SSIM (MS-SSIM)* is an extension of the SSIM paradigm, also proposed for still images,[18] that has been shown to outperform the SSIM index and many other still image quality assessment algorithms.[19] 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 freely available for download.[17]
- *Speed SSIM* is the name we give to the VQA model[20] that uses the SSIM index in conjunction with statistical models of visual speed perception.[21] Using models of visual speed perception was shown to improve the performance of both PSNR and SSIM in.[20] We evaluated the performance of this framework with the SSIM index, which was shown to perform better than using the same framework with PSNR.[20] 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[22] and is freely available for download.[23] 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).[24] 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 for research purposes.[25]
- *V-VIF* is the name we give to the VQA model[26] that extends the Visual Information Fidelity (VIF) criterion for still images[27] 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.[28,29] Three different versions of the MOVIE index - the Spatial MOVIE index, the Temporal MOVIE index and the MOVIE index - were tested in our study.

We tested the performance of all objective models using two metrics-the Spearman Rank Order Correlation Coefficient (SROCC) which measures the monotonicity of the VQA algorithm prediction against human scores and the Pearson Linear Correlation Coefficient (LCC) which measures prediction accuracy. The LCC is computed after performing a non-linear regression on the VQA algorithm scores to map them to DMOS scores using a logistic function.[2] 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.[2]

$$Q'_j = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-\left(\frac{Q_j - \beta_3}{|\beta_4|}\right)}} \tag{4}$$

Prediction Model	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

Table 1: Comparison of the performance of VQA algorithms - SROCC. The best performing objective VQA algorithm is highlighted in bold font for each category.

Prediction Model	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

Table 2: Comparison of the performance of VQA algorithms - LCC. The best performing objective VQA algorithm is highlighted in bold font for each category.

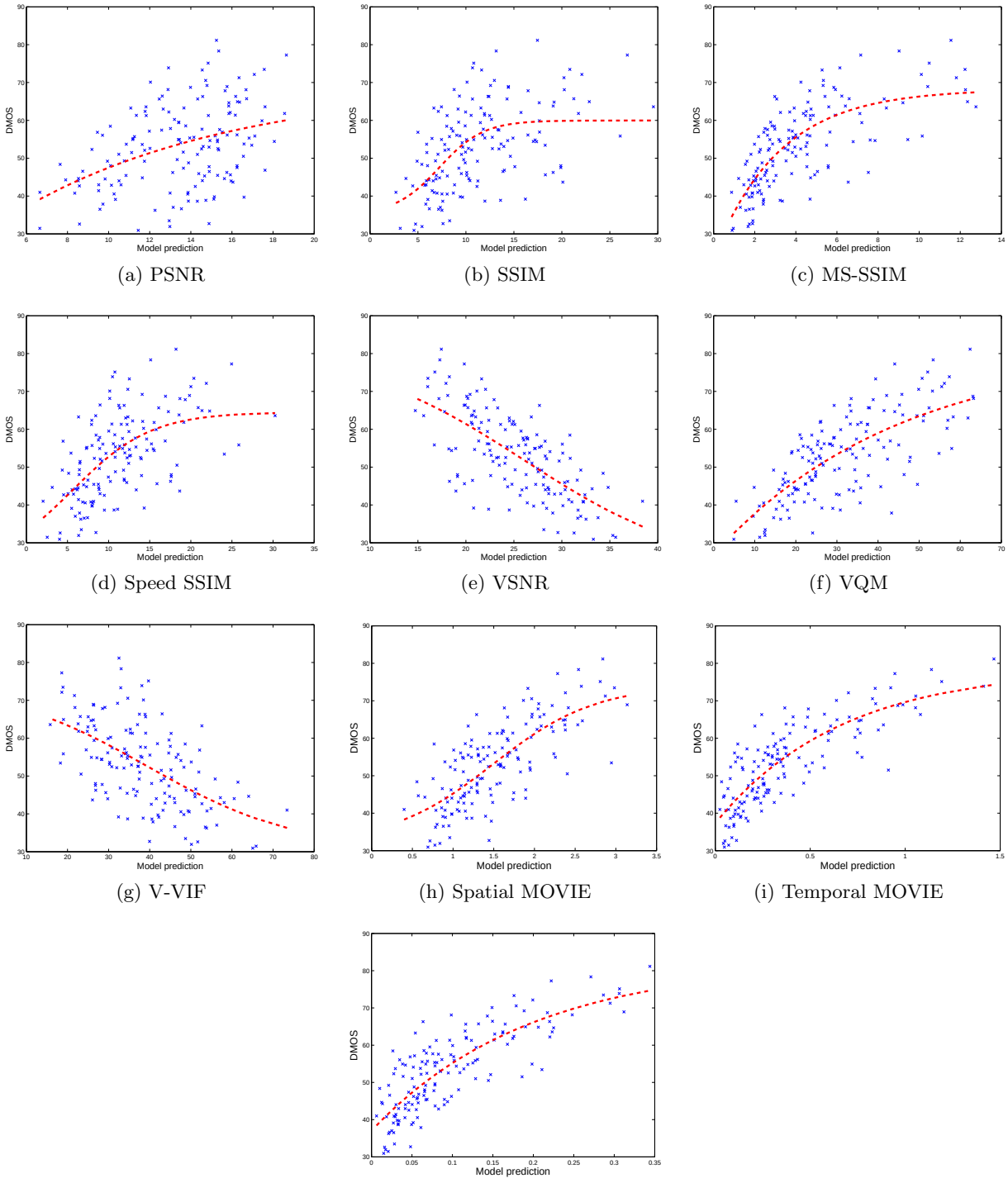
Non-linear least squares optimization is performed using the Matlab function "nlinfit" to find the optimal parameters β that minimize the least squares error between the vector of subjective scores ($\text{DMOS}_j, j = 1, 2, \ldots 150$) and the vector of fitted objective scores ($Q'_j, j = 1, 2, \ldots, 150$). Initial estimates of the parameters were chosen based on the VQEG recommendation.[2] 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).

Tables 1 and 2 show the performance of all models in terms of the SROCC and LCC separately for each distortion type and for the entire database. Scatter plots of objective scores vs. DMOS for all the algorithms on the entire LIVE Video Quality Database, along with the best fitting logistic functions, are shown in Fig. 3.

4. DISCUSSION

Our results clearly demonstrate that a carefully constructed database of videos can expose the significant limitations of PSNR as a VQA measure. PSNR is shown to perform very poorly in correlating with human subjective judgments and is clearly an unreliable predictor of quality in any application where the end user of the video is a human observer.

All the perceptual VQA algorithms tested in our study improve upon PSNR. The MS-SSIM index improves upon the Frame SSIM index. The Speed SSIM index utilizes the Frame SSIM index to compute local quality estimates and uses models of visual speed perception to weight these local estimates and pool them into an overall quality estimate for the video. Speed SSIM improves over the Frame SSIM index, although it does not perform as well as the MS-SSIM index. The MS-SSIM index[18] and the VQM from NTIA[24] perform quite well on the LIVE Video Quality Database and are well-suited for video benchmarking applications where low-complexity



(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

Figure 3: Scatter plots of objective VQA scores vs. DMOS for all videos in the LIVE Video Quality Database. Also shown is the best fitting logistic function.

and fast VQA algorithms are desired, since both algorithms do not perform computationally intensive operations such as motion estimation.

The best performing VQA algorithm amongst the ones tested in our study, in terms of both the SROCC and LCC after non-linear regression, is the Temporal MOVIE index. One of the three versions of the MOVIE index (Spatial MOVIE, 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. 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.[28,29] 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.[28,29] 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. MOVIE utilizes specific (Gabor receptive field) models of cortical area V1 to dissemble video data into multi-scale space-time primitives. MOVIE also uses a specific model of the relatively well-understood extra-cortical area V5 (also known as Area MT) to effect a biologically plausible model of visual motion processing.[30] Using these models, MOVIE deploys SSIM-like multi-scale processing to compute local scale-space comparisons that can be supported from an information-theoretic viewpoint under natural scene statistical models.[31]

Looking at the break-down 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. This is unsurprising since Spatial MOVIE performs a multi-scale decomposition similar to MS-SSIM, but one that is perceptually matched 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, despite not being tuned to detect spatial distortions (of which the database contains many). Temporal MOVIE is unique amongst all algorithms tested in our study since it models visual motion perception. MOVIE also shows excellent performance. It is interesting that the performance of Temporal MOVIE is better than that of MOVIE overall. However, MOVIE performs better than Temporal MOVIE on the wireless and IP videos (in terms of LCC) and on the VQEG database.[29] The success of MOVIE and the improved performance of Speed SSIM over Frame SSIM strongly validates the notion that using computed motion information can improve VQA algorithm performance.

5. CONCLUSIONS

A subjective study to evaluate the effects of present generation video technology 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. We presented an evaluation of the performance of several publicly available objective VQA models on this database and the MOVIE index performed the best amongst the algorithms we tested. In the future, we would like to study different spatial and temporal pooling strategies for VQA, which is particularly relevant in the context of spatially and temporally localized errors in videos that often occur in video communication systems.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Science Foundation (Award Number: 0728748).

REFERENCES

[1] Seshadrinathan, K., Soundararajan, R., Bovik, A. C. and Cormack, L. K., "LIVE Video Quality Database." http://live.ece.utexas.edu/research/quality/live_video.html (2009).

[2] The Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment." http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI (2000).

[3] Sheikh, H. R. and Bovik, A. C., "LIVE image quality assessment database." http://live.ece.utexas.edu/research/quality/subjective.htm (2003).

[4] Technical University of Munich, "High defininiton videos." ftp://ftp.ldv.e-technik.tu-muenchen.de/pub/test_sequences/ (2003).

[5] International Organization for Standardization, "MPEG-2 standard." http://standards.iso.org/ittf/PubliclyAvailableStandards/ (2005).
[6] Joint Video Team, "H.264/AVC software coordination." http://iphome.hhi.de/suehring/tml/ (2007).
[7] Wenger, S., "H.264/AVC over IP," *IEEE Transactions on Circuits and Systems for Video Technology* **13**, 645–656 (July 2003).
[8] Video Coding Experts Group, "Proposed error patterns for internet experiments." http://ftp3.itu.ch/av-arch/video-site/9910_Red/q15i16.zip (1999).
[9] Stockhammer, T., Hannuksela, M. M., and Wiegand, T., "H.264/AVC in wireless environments," *IEEE Transactions on Circuits and Systems for Video Technology* **13**, 657–673 (July 2003).
[10] Video Coding Experts Group, "Common Test Conditions for RTP/IP over 3GPP/3GPP2." http://ftp3.itu.ch/av-arch/video-site/0109_San/VCEG-N80_software.zip (1999).
[11] Perry, J., "The XGL Toolbox." http://128.83.207.86/∼jsp/software/xgltoolbox-1.0.5.zip (2008).
[12] van Dijk, A. M., Martens, J.-B., and Watson, A. B., "Quality asessment of coded images using numerical category scaling," in [*Proc. SPIE - Advanced Image and Video Communications and Storage Technologies*], (1995).
[13] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," tech. rep., International Telecommunications Union (2000).
[14] Wang, Z. and Bovik, A. C., "A universal image quality index," *IEEE Signal Processing Letters* **9**(3), 81–84 (2002).
[15] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004).
[16] Wang, Z., Lu, L., and Bovik, A. C., "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication* **19**, 121–132 (Feb. 2004).
[17] Wang, Z., Sheikh, H. R., Bovik, A. C., and Simoncelli, E. P., "Image and video quality assessment at LIVE." http://live.ece.utexas.edu/research/Quality/index.htm (2004).
[18] Wang, Z., Simoncelli, E. P. ., Bovik, A. C., and Matthews, M. B., "Multiscale structural similarity for image quality assessment," in [*IEEE Asilomar Conf. Signals, Sys. and Comp.*], (2003).
[19] Sheikh, H. R. and Bovik, A. C., "An evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing* **15**, 3440–3451 (November 2006).
[20] Wang, Z. and Li, Q., "Video quality assessment using a statistical model of human visual speed perception.," *Journal of the Optical Society of America A - Optics, Image Science and Vision* **24**, B61–B69 (Dec 2007).
[21] Stocker, A. A. and Simoncelli, E. P., "Noise characteristics and prior expectations in human visual speed perception.," *Nature Neuroscience* **9**, 578–585 (Apr 2006).
[22] Chandler, D. M. and Hemami, S. S., "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing* **16**(9), 2284–2298 (2007).
[23] Chandler, D. M. and Hemami, S. S., "MeTriX MuX Visual Quality Assessment Package." http://foulard.ece.cornell.edu/gaubatz/metrix_mux/ (2007).
[24] Pinson, M. H. and Wolf, S., "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting* **50**, 312–322 (Sept. 2004).
[25] NTIA, "VQM." http://www.its.bldrdoc.gov/n3/video/VQM_software.php (2008).
[26] Sheikh, H. R. and Bovik, A. C., "A visual information fidelity approach to video quality assessment," in [*First International conference on video processing and quality metrics for consumer electronics*], (2005).
[27] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," *IEEE Transactions on Image Processing* **15**(2), 430–444 (2006).
[28] Seshadrinathan, K. and Bovik, A. C., "Motion-based perceptual quality assessment of video," in [*Proc. SPIE - Human Vision and Electronic Imaging*], (2009).
[29] Seshadrinathan, K. and Bovik, A. C., "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing* **2**(9) (In press, Feb. 2010.).
[30] Simoncelli, E. P. and Heeger, D. J., "A model of neuronal responses in visual area MT," *Vision Research* **38**, 743–761 (Mar 1998).
[31] Seshadrinathan, K. and Bovik, A. C., "Unifying analysis of full reference image quality assessment," in [*IEEE Intl. Conf. on Image Proc.*], (2008).