SUBJECTIVE ANALYSIS OF VIDEO QUALITY ON MOBILE DEVICES

Anush K. Moorthy, Lark K. Choi, Gustavo de Veciana and Alan C. Bovik

Department of Electrical and Computer Engineering,
The University of Texas at Austin,
Austin, Tx - 78712, USA.

ABSTRACT

We summarize a recent mobile video quality assessment (VQA) database that we have created which incorporates a wide variety of distortions including uniform compression and wireless packet loss, along with dynamically varying distortions that change as a user views a video on his mobile device. We evaluate a large number of objective image and video quality assessment (IQA/VQA) algorithms with regards to their efficacy in predicting visual quality on these videos on such small form-factor displays. A detailed correlation analysis and statistical hypothesis testing is carried out. Our results indicate that current algorithms for IQA and VQA are not well-equipped to handle time-varying distortions and need to be augmented substantially to take into account varied viewing conditions and form-factors of devices.

1. INTRODUCTION

The CISCO Visual Networking Index (VNI) global mobile data traffic forecast states that mobile video traffic accounts for nearly 50% of mobile traffic, and by 2015, this percentage will steadily increase to more than 75% [1]. With the recent explosion of smart phones and tablets, and the emergence of video streaming websites and services such as Amazon Video on Demand, Hulu, Netflix, YouTube etc., mobile video traffic is fast becoming a burden on the limited amount of spectrum. The paucity of bandwidth is evident from the bandwidth caps that most of the wireless providers in the U.S. have recently imposed on data-hungry users.

With this burgeoning demand for video content, it is imperative that researchers develop frameworks for wireless networks that are capable of handling video traffic in an efficient manner. One promising direction is the *perceptual optimization* of wireless video networks, wherein network resource allocation protocols are designed to provide video experiences that are measurably improved under perceptual models. Since the final receivers of videos are almost always human observers, visual perception is the ultimate arbiter of the visual experience. In order to understand and hence model human opinion on visual quality, we have created a large-scale database of HD videos that humans viewed and rated for their quality on mobile devices.

There have been several subjective studies that have been conducted in the past with various aims [2–6]. These subjective studies have been performed on large format displays and the distorted videos have typically included compressed videos, videos transmitted over wireless/IP networks [2, 3], and delayed and jittered videos [7, 8]. While these databases have definite value in understanding human perception of quality and the development of objective quality assessment algorithms, the results obtained do not translate to smaller screen resolutions.

Some video quality studies have also been conducted on mobile devices [9–13]. However, almost all of these studies suffer from several limitations: (1) small datasets, (2) insufficient distortions and distortion severities, (3) unknown sources for the reference videos used with unknown corruptions, (4) small video resolutions and (5) lack of public availability.

The LIVE Mobile VQA database that we describe here aims to overcome the above limitations. This adequate and modern database will aid the development of perceptually optimized algorithms for wireless video transmission and serve as a testing ground for VQA algorithms. The LIVE Mobile VQA database consists of 200 distorted videos evaluated by over 30 human subjects on a small mobile screen. The RAW HD videos used in the study were captured using a RED ONE digital cinematographic camera and the distortions include compression and wireless channel transmission losses. More importantly, the LIVE mobile VQA database also includes dynamically changing distortions resulting in perceptible quality fluctuations in the videos over time. Summary subjective scores at the end of each presentation and continuously recorded scores were collected. We also evaluated the performance of a wide range of image and video quality assessment (IQA/VQA) algorithms in terms of correlation with human perception for each distortion and across distortion categories. The LIVE Mobile VQA database is being made available online in order to help further research in the area of visual quality assessment.

This article summarizes the construction of the LIVE Mobile VQA database and the associated subjective study. The results of evaluating leading full-reference (FR) image/video quality assessment (I/VQA) algorithms for their ability to predict visual quality, including hypothesis testing and statistical significance analysis, is also detailed. We believe that the LIVE Mobile VQA database will be an invaluable tool for research in video quality assessment and the perceptual optimization of wireless video delivery.

2. SUMMARY OF SUBJECTIVE STUDY

2.1. Source Videos and Distortion Simulation

A digital cinematographic camera – the RED ONE – was used to capture 12-bit REDCODE RAW data at a resolution of $2K(2048 \times 1152)$ at frame rates of 30 fps and 60 fps using the REDCODE 42MB/s option to obtain the best possible quality. Source videos were downsampled to a resolution of 1280×720 ($720p$) at 30 and 15 second portions of the videos were selected so that the video content spans a wide range of scenes and lighting conditions. Twelve such videos were converted into uncompressed .yuv files of which two videos were used to train the subjects, and the remaining 10 videos were used in the actual study (see below). Figure 1 shows sample frames from some of the videos.

Fig. 1. Example frames of the videos used in the study.

We simulated uniform distortions such as compression and wireless packet-loss as well as dynamic distortions such as frame-freezes, rate adaptation and temporal dynamics. We summarize the distortion simulation below.

The JM reference implementation of the H.264 scalable video codec (SVC) [14, 15] was used to compress each reference video at four compression rates (R_1, R_2, R_3, R_4; $R_1 < R_2 < R_3 < R_4$) between 0.7 Mbps and 6 Mbps using fixed QP encoding. The QP values were manually selected as in [2], to ensure perceptual separation between the distorted videos for each content. Each compressed video was transmitted through a simulated Rayleigh fading channel. The simulated channel was modeled using an IEEE 802.11- based wireless channel simulator implemented in LabVIEW. The system comprised of a single link channel with coding, interleaving, QAM modulation, and OFDM modulation. A large set of error traces were first generated using the simulated channel and a randomly chosen error trace were then applied to each packetized compressed video.

We simulated two frame-freeze models to model frame-freezes induced during: (1) stored video delivery and (2) live video delivery. For stored video delivery, a frame-freeze did not result in the loss of a video segment, i.e., the video played back from the same temporal location after the freeze. In the case of live video delivery, the video segment corresponding to the length of the freeze was lost and hence the real-time frame-freezes lacked temporal continuity. We simulated three stored video freeze lengths of 1 second, 2 seconds and 4 seconds, and one real-time video freeze length of 4 seconds. In all cases, there was no freeze in the first 3 seconds of playback and the video duration after the freeze lasted 1.5× the length of the freeze before the next freeze occurred.

The rate adaptation condition simulated rate changes in the video as a function of time. The videos starts out at rate R_X, then after n seconds switches to a higher rate R_Y, then again after n seconds switches back to the original rate R_X. Three different rate switches were simulated, where n was fixed at 5 seconds, so that $R_X = R_1, R_2$ and R_3 and $R_Y = R_4$, as illustrated in Fig. 2.

The temporal dynamics was simulated to model the effect of multiple rate switches and to evaluate the effect of the *abruptness* of the switch. Five different conditions were simulated. In the first condition, the rate was varied between R_1 and R_4 multiple times (3), as illustrated in Fig. 3. The remaining four conditions were as follows: (1) $R_1 - R_2 - R_4$, (2) $R_1 - R_3 - R_4$, (3) $R_4 - R_2 - R_1$ and (4) $R_4 - R_3 - R_1$, as illustrated in Fig. 4. While the first condition models multiple rate switches, the remaining four conditions were

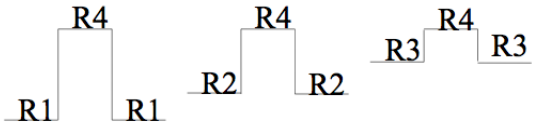


Fig. 2. Rate Adaptation: Schematic diagram of the three different rate-switches in a video stream simulated in this study.

Fig. 3. Temporal Dynamics: Schematic illustration of two rate changes across the video; the average rate remains the same in both cases. Left: Multiple changes and Right: Single rate change. Note that we have already simulated the single rate-change condition as illustrated in Fig. 2, hence we ensure that the average bit-rate is the same for these two cases.

modeled to reveal whether easing a user into a higher/lower quality regime is better than abruptly switching between these two regimes. Note the dual structures of the conditions. These structures were chosen to answer the question: Is ending a video with a high quality segment better than ending it on a low-quality one.

In summary, the LIVE Mobile VQA database consists of 10 reference videos and 200 distorted videos (4 compression + 4 wireless packet-loss + 4 frame-freezes + 3 rate-adapted + 5 temporal dynamics per reference), each of resolution 1280×720 at a frame rate of 30 fps, and of duration 15 seconds each.

2.2. Test Methodology

Students from The University of Texas at Austin participated in the three-week long single-stimulus continuous quality evaluation (SS-CQE) study [16] with hidden reference [2, 3, 17] at the LIVE subjective testing lab, where they viewed the videos on the Motorola Atrix. The dual-core, 4" device has a display resolution of 960×540, and is capable of displaying $720p$ videos smoothly. An Android app was written specifically for this purpose, which displayed the videos at the center of the screen with a continuous uncalibrated rating bar,

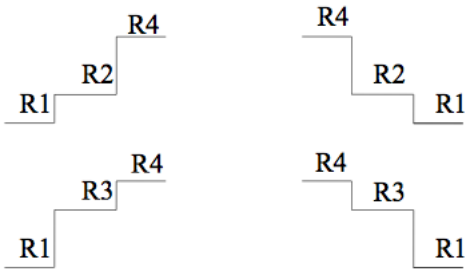


Fig. 4. Temporal Dynamics: Schematic illustration of rate-changes scenarios. The average rate remains the same in all cases and is the same as in Fig. 3. The first row steps to rate R_2 and then steps to a higher/lower rate, while the second row steps to R_3 and then back up/down again

controlled by using the touchscreen. The subject rated the videos as s/he watched the presentations, yielding continuous temporal quality scores. At the end of each presentation, a similar rating bar, albeit with calibration (Bad-Fair-Excellent) was presented to the subject, where s/he now provided an overall quality score for the video. Fig. 5 shows the various stages of the study.

The study involved mostly (naive) undergraduate students whose average age was between 22-28 years. While a verbal confirmation of soundness of (corrected) vision was obtained from the subject, no vision test was performed. This was in keeping with our philosophy of using a reasonably representative sampling of the visual population. Each subject attended two separate sessions as part of the study such that each session lasted less than 30 minutes, each of which consisted of the subject viewing (randomized) 55 videos (50 distorted + 5 reference); a short training set (6 videos) preceded the actual study.

Thirty-six subjects participated in the study, and the design of the study was such that each video received 18 subjects. Two subjects were rejected using the procedure outlined in [16]. Since this was a hidden-reference study, the reference pristine videos were included in the set of the videos that the subject saw, and ratings were collected from the subject for these pristine videos. A differential mean opinion score (DMOS) was computed as the difference between the score that the subject gave the reference and the score for the distorted videos (after subject rejection). DMOS computation was performed only for the final summary scores, and we analyze only these summary DMOS scores in this article.

DMOS values ideally range continuously from 0 (excellent quality) to 5 (worst quality); however small negative values as possible due to the nature of DMOS computation. Figure 6 plots the DMOS scores across distorted videos, and shows the corresponding histogram in order to demonstrate that the distorted videos span the entire quality range. The average standard error in the DMOS score was 0.2577 across the 200 distorted videos. We assume that the DMOS scores sample a Gaussian distribution centered around the DMOS having a standard deviation computed from the differential opinion scores across subjects for all further analysis.

3. EVALUATION OF ALGORITHM PERFORMANCE

A variety of full-reference (FR) image quality assessment (IQA) algorithms, and some FR video quality assessment (VQA) algorithms were evaluated against the collected human subjective opinion scores. All evaluations were performed only on the luminance images. The FR IQA algorithms evaluated in this study are listed in Table 1, all of which are available as part of the Metrix Mux toolbox [18]. The reader is referred to the citations for details on these approaches. FR IQA algorithms were applied on a frame-by-frame basis and the final scores obtained for the video was the time-average of the frame-level quality scores. Since it is unclear how FR QA algorithms may be used for frame-freezes, we did not include this case in our evaluations.

No.	Algorithm
1.	Peak Signal-to-Noise ratio (PSNR)
2.	Structural Similarity Index (SS-SSIM) [19]
3.	Multi-scale Structural Similarity Index (MS-SSIM) [20]
4.	Visual Signal-to-Noise ratio (VSNR) [21]
5.	Visual Information Fidelity (VIF) [22]
6.	Universal Quality Index (UQI) [23]
7.	Noise Quality Measure (NQM) [24]
8.	Signal-to-Noise ratio (SNR)
9.	Weighted Signal-to-Noise ratio (WSNR) [25]

Table 1. List of FR 2D IQA algorithms evaluated in this study.

The FR VQA algorithms evaluated in this study were the Visual Quality Metric (VQM) [26] and the MOtion-based Video Integrity Evaluation (MOVIE) index [27]. VQM was obtained from [28] while MOVIE is freely available at [29]. The version of VQM that we used (CVQM v13) requires input videos in YUV422p format encased in an avi container. The YUV420p videos were converted to YUV422p using ffmpeg, then placed in an avi container (no compression was used). These algorithms were also not evaluated for their performance on frame-freezes.

Table 2 tabulates the Spearman's rank ordered correlation coefficient (SROCC), Table 3 tabulates the Pearson's (linear) correlation coefficient (LCC) and Table 4 tabulates the root mean-squared-error (RMSE) between the algorithm scores and DMOS. LCC and RMSE were computed after non-linear regression using a logistic function prescribed in [30][1]). The SROCC, LCC and RMSE values are listed for each distortion-class, as well as across all simulated distortions. The best performing algorithm (SROCC/LCC) in each column is highlighted.

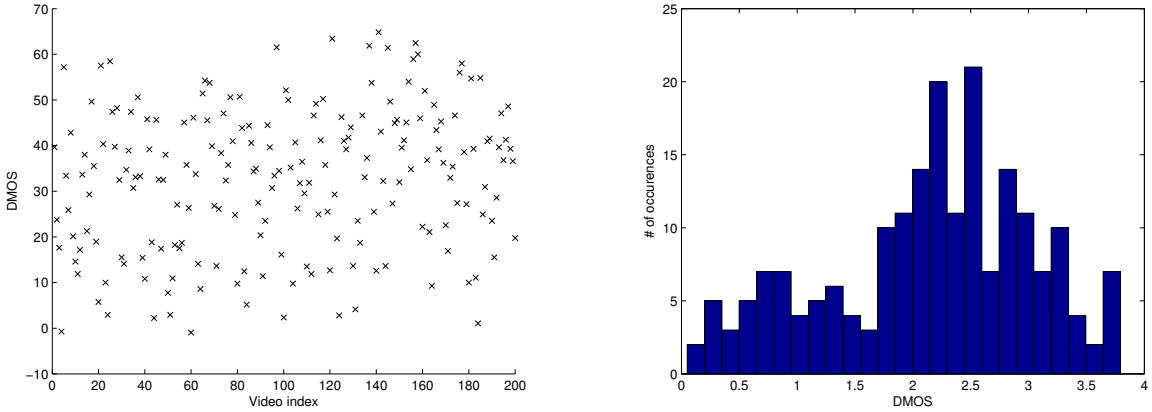
The results of the study support two immediate conjectures. First, the tables indicate that multiscale processing matters with a reduction in display size. The two truly multiscale algorithms,VSNR and VIF, both of which use wavelet decompositions yield the best overall performance, exceeding that of true VQA algorithms – the single-scale VQM and the partially-multiscale MOVIE index, which omits high frequencies. While MS-SSIM does well, its weighting scheme, which over weights the mid-band frequencies, possibly reduces its overall performance. A lesson here is that true multiscale is advisable to achieve scalability against variations in display size, resolution and viewing distance, suggesting future refinements of VQA algorithms.

Second, almost all algorithms fail to reliably predict overall subjective judgements of dynamic distortions, especially for the set of "temporal-dynamics" distorted videos. Some algorithms perform well for the wireless and compression distortions – e.g., VQM, NQM and VIF. While the algorithm performance for the rate-adaptation case demonstrates that there remains significant room for improve-

[1]Except for MOVIE, where the fitting failed; instead the logistic specified in [5] was used.

(a) (b)

Fig. 5. Study Setup: (a) The video is shown at the center of the screen and an (uncalibrated) bar at the bottom is provided to rate the videos as a function of time. The rating is controlled using the touchscreen. (b) At the end of the presentation, a similar calibrated bar is shown on the screen so that the subject may rate the overall quality of the video.



(a) (b)

Fig. 6. (a) Scatter plot of DMOS for distorted videos. (b) Histogram of DMOS obtained from the study.

	Comp.	RA	TD	WL	All
PSNR	0.8185	0.5981	0.3717	0.7925	0.6780
SS-SSIM	0.7092	0.6303	0.3429	0.7246	0.6498
MS-SSIM	0.8044	**0.7378**	**0.3974**	0.8128	0.7425
VSNR	**0.8739**	0.6735	0.3170	0.8559	**0.7517**
VIF	0.8613	0.6388	0.1242	0.8739	0.7439
UQI	0.5621	0.4299	0.0296	0.5756	0.4894
NQM	0.8499	0.6775	0.2383	**0.8985**	0.7493
WSNR	0.7817	0.5598	0.0942	0.7510	0.6267
SNR	0.7073	0.5565	0.2029	0.6959	0.5836
VQM	0.7717	0.6475	0.3860	0.7758	0.6945
MOVIE	0.7738	0.7198	0.1578	0.6508	0.6420

Table 2. Mobile Study: Spearman's Rank ordered correlation coefficient (SROCC) between the algorithm scores and the DMOS for various IQA/VQA algorithms. Comp = compression, RA = rate adaptation, TD = temporal dynamics, WL = wireless loss.

	Comp.	RA	TD	WL	All
PSNR	0.7841	0.5364	0.4166	0.7617	0.6909
SS-SSIM	0.7475	0.6120	0.3924	0.7307	0.6637
MS-SSIM	0.7664	0.7089	0.4068	0.7706	0.7077
VSNR	0.8489	0.6581	**0.4269**	0.8493	0.7592
VIF	**0.8826**	0.6643	0.1046	**0.8979**	**0.7870**
UQI	0.5794	0.2929	0.2546	0.7412	0.6619
NQM	0.8318	0.6772	0.3646	0.8738	0.7622
WSNR	0.7558	0.5365	0.0451	0.7276	0.6320
SNR	0.6501	0.3988	0.0839	0.6052	0.5189
VQM	0.7816	0.5910	0.4066	0.7909	0.7023
MOVIE	0.8103	**0.6811**	0.2436	0.7266	0.7157

Table 3. Mobile Study: Linear (Pearson's) correlation coefficient (LCC) between the algorithm scores and the DMOS for various IQA/VQA algorithms. Comp = compression, RA = rate adaptation, TD = temporal dynamics, WL = wireless loss.

ment, MS-SSIM and MOVIE perform reasonably well. Overall, VSNR, VIF, MS-SSIM and NQM are seemingly well correlated with human perception, while the single-scale UQI, which captures the narrowest range of frequencies, is the weakest of the lot. The widely criticized PSNR holds its own against compression and wireless distortions, since, while it is not multiscale, it captures high frequency

	Comp.	RA	TD	WL	All
PSNR	0.7069	0.5733	0.4179	0.7279	0.6670
SS-SSIM	0.7566	0.6023	0.4228	0.7670	0.6901
MS-SSIM	0.7316	0.4792	0.4199	0.7160	0.6518
VSNR	0.6021	0.5115	0.4157	0.5932	0.6005
VIF	0.5354	0.5078	0.4572	0.4945	0.5692
UQI	0.9283	0.6496	0.4445	0.7542	0.6916
NQM	0.6374	0.4999	0.4280	0.5463	0.5972
WSNR	0.7458	0.5733	0.4592	0.7707	0.7150
SNR	0.8654	0.6230	0.4580	0.8944	0.7887
VQM	0.7312	0.4840	0.4141	0.7279	0.6663
MOVIE	0.6674	0.4974	0.4458	0.7719	0.6444

Table 4. Mobile Study: Root mean-squared-error (RMSE) between the algorithm scores and the DMOS for various IQA/VQA algorithms. Comp = compression, RA = rate adaptation, TD = temporal dynamics, WL = wireless loss.

distortions.

To gauge if the correlations were significantly different from each other, we performed a statistical analysis of the algorithm scores using the F-statistic as in [2, 30]. Specifically, the F-statistic was used to evaluate the difference between the variances of the residuals produced after a non-linear mapping between the two algorithms being compared, and Table 5 lists the results for each distortion category and across all distortions. A value of '1' in the tables indicates that the row (algorithm) is statistically better than the column (algorithm), while a value of '0' indicates that the row is worse than the column; a value of '-' indicates that the row and column are statistically identical.

Tables 5 validates our observations from the correlations – NQM, VIF, VQM perform well, although interestingly, NQM is the only algorithm that is statistically superior to PSNR overall for the study.

4. DISCUSSION AND CONCLUSION

We summarized a new resource – the LIVE Mobile video quality assessment (VQA) database – that incorporates 200 distorted videos and associated human opinion scores of video quality from over 30 subjects which is being made available to researchers at no cost. The distortion simulated included the previously studied compression and wireless loss artifacts, and novel dynamically varying distortions such as rate changes and frame-freezes. We analyzed various leading FR VQA algorithms for their ability to predict video quality on mobile devices. Our results indicate that none of the present-day algorithms are capable of predicting video quality accurately for the dynamic distortions. More importantly, our results lead to the observation that multiscale processing is of importance when assessing quality on devices with varying resolutions. The LIVE Mobile VQA database will be a fertile testing ground for future algorithm development. We have only summarized a portion of the database, and the reader is referred to [31] for a thorough analysis of the human opinion scores and a detailed description of the study.

5. REFERENCES

[1] CISCO Corp., "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010–2015," *http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html*, 2011.

[2] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 1427–1441, 2010.

[3] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms'," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 513–516, April 2010.

[4] Video Quality Experts Group (VQEG), "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment phase II," *http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII*, 2003.

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

[6] Video Quality Experts Group (VQEG), "Final report of video quality experts group multimedia phase I validation test, TD 923, ITU Study Group 9," 2008.

[7] S.R. Gulliver and G. Ghinea, "The perceptual and attentive impact of delay and jitter in multimedia delivery," *Broadcasting, IEEE Transactions on*, vol. 53, no. 2, pp. 449–458, 2007.

[8] Q. Huynh-Thu and M. Ghanbari, "Impact of jitter and jerkiness on perceived video quality," in *Proceedings of the Workshop on Video Processing and Quality Metrics*. Citeseer, 2006.

[9] A. Eichhorn and P. Ni, "Pick your layers wisely-a quality assessment of h. 264 scalable video coding for mobile devices," *Proceedings of the 2009 IEEE international conference on Communications*, pp. 5446–5451, 2009.

[10] H. Knoche, J.D. McCarthy, and M.A. Sasse, "Can small be beautiful?: assessing image resolution requirements for mobile TV," *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 829–838, 2005.

[11] S. Jumisko-Pyykko and J. Hakkinen, "Evaluation of subjective video quality of mobile devices," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 535–538.

[12] M. Ries, O. Nemethova, and M. Rupp, "Performance evaluation of mobile video quality estimators," in *Proceedings of the European Signal Processing Conference,(Poznan, Poland*. Citeseer, 2007.

[13] S. Jumisko-Pyykko and M.M. Hannuksela, "Does context matter in quality evaluation of mobile television?," in *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*. ACM, 2008, pp. 63–72.

[14] Joint Video Team (JVT), "SVC Reference Software (JSVM software)," *http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm*.

[15] Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, "Joint Draft ITU-T Rec. H.264 — ISO/IEC 14496-10 / Amd.3 Scalable video coding," *http://www.hhi.fraunhofer.de/fileadmin/hhi/downloads/IP/ip_ic_H.264-MPEG4-AVC-Version8-FinalDraft.pdf*.

	PSNR	SS-SSIM	MS-SSIM	VSNR	VIF	UQI	NQM	WSNR	SNR	VQM	MOVIE
PSNR	- - - - -	1 - - 1 1	- 0 - - -	- - - - -	- 1 1 - -	1 1 1 1 1	- - - 0 0	1 1 1 1 1	1 1 1 1 1	1 0 - - -	1 - 1 1 -
SS-SSIM	0 - - 0 0	- - - - -	0 0 - 0 0	0 - - 0 0	0 1 1 0 0	- 1 1 - 1	0 - - 0 0	- 1 1 - -	- - - - 1	- 0 - 0 0	- - 1 - -
MS-SSIM	- 1 - - -	1 1 - 1 1	- - - - -	- 1 - - -	- 1 - - -	1 1 - 1 1	0 - - 0 0	- 1 - - 1	1 1 - 1 1	- - - - -	- - - - 1
VSNR	- - - - -	1 - - 1 1	- 0 - - -	- - - - -	- 1 1 - -	1 1 1 1 1	- - - 0 -	1 1 1 1 1	1 1 1 1 1	1 0 - - -	1 - 1 1 1
VIF	- 0 0 - -	1 0 0 1 1	- 0 - - -	- 0 0 - -	- - - - -	1 - - 1 1	- 0 0 0 0	1 - - 1 1	1 - - 1 1	1 0 0 1 -	1 0 - 1 -
UQI	0 0 0 0 0	- 0 0 - 0	0 0 - 0 0	0 0 0 0 0	0 - - 0 0	- - - - -	0 0 - 0 0	0 - - - 0	- - - - -	0 0 - 0 0	0 0 - - 0
NQM	- - - 1 1	1 - - 1 1	1 - - 1 1	- - - 1 -	- 1 1 1 1	1 1 - 1 1	- - - - -	1 1 - 1 1	1 1 - 1 1	1 0 - 1 1	1 - 1 1 1
WSNR	0 0 0 0 0	- 0 0 - -	- 0 - - 0	0 0 0 0 0	0 - - 0 0	1 - - - 1	0 0 - 0 0	- - - - -	- - - - -	- 0 0 - 0	- 0 - - 0
SNR	0 0 0 0 0	- - - - 0	0 0 - 0 0	0 0 0 0 0	0 - - 0 0	- - - - -	0 0 - 0 0	- - - - -	- - - - -	- 0 - 0 0	- 0 - - 0
VQM	0 1 - - -	- 1 - 1 1	- - - - -	0 1 - - -	0 1 1 0 -	1 1 - 1 1	0 1 - 0 0	- 1 1 - 1	- 1 - 1 1	- - - - -	- - 1 - -
MOVIE	0 - 0 0 -	- - 0 - -	- - - - 0	0 - 0 0 0	0 1 - 0 -	1 1 - - 1	0 - 0 0 0	- 1 - - 1	- 1 - - 1	- - 0 - -	- - - - -

Table 5. Mobile Study: Statistical analysis of algorithm performance. A value of '1' in the tables indicates that the row (algorithm) is statistically better than the column (algorithm), while a value of '0' indicates that the row is worse than the column; a value of '-' indicates that the row and column are statistically identical. Within each entry of the matrix, the first four symbols correspond to the four distortions (ordered as in the text), and the last symbol represents significance across the entire database.

[16] International Telecommunication Union (ITU), "BT-500-11: Methodology for the subjective assessment of the quality of television pictures," *Recommendation*.

[17] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *Visual Communications and Image Processing, SPIE*, vol. 5150, 2003.

[18] Matthew Gaubatz, "Metrix mux visual quality assessment package," *http://foulard.ece.cornell.edu/gaubatz/metrix_mux/*.

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

[20] Z. Wang, L. Lu, and A.C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Transactions on Image Processing*, vol. 12, no. 2, pp. 243, 2003.

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

[22] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[23] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[24] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, 2002.

[25] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 525–535, 1974.

[26] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

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

[28] "Video quality metric," *http://www.its.bldrdoc.gov/n3/video/VQM_software.php*.

[29] K. Seshadrinathan, "Movie software release," *http://live.ece.utexas.edu/research/Quality/movie.html*, 2010.

[30] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[31] A. K. Moorthy, L-K. Choi, A. C. Bovik, and G. de Veciana, "Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on New Subjective and Objective Methodologies for Audio and Visual Signal Processing*, 2011 (submitted).