Copyright © 2012 IEEE.
Citation for the published paper:

Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos

Muhammad Shahid, Amitesh Kumar Singam, Andreas Rossholm, Benny Lövstrom

5th International Congress on Image and Signal Processing (CISP)

2012 Chongqing, Sichuan, China

This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of BTH's products or services Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by sending a blank email message to pubs-permissions@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos

Muhammad Shahid, Amitesh Kumar Singam, Andreas Rossholm and Benny Lövström
Department of Signal Processing, Blekinge Institute of Technology
SE-37179 Karlskrona, Sweden
Corresponding author email: muhammad.shahid@ieee.org

***Abstract*—Advancements in the video processing area have been proliferated by services that require low delay. Such services involve applications being offered at various temporal and spatial resolutions. It necessitates to study the impacts of related video coding conditions upon perceptual quality. But most of studies concerned with quality assessment of videos affected by coding distortions lack in variety of spatio-temporal resolutions. This paper presents a work done on quality assessment of videos encoded by state-of-the-art H.264/AVC standard at different bitrates and frame rates. Overall, 120 test scenarios for video sequences having different spatial and temporal spectral information were studied. The used coded bistreams in this work and the corresponding subjective assessment scores have been made public for the research community to facilitate further studies.**

I. Introduction

The recent advancement in digital imaging technology and availability of efficient transmission systems have resulted in a proliferation of videos more than ever before. Videos transmitted to and from mobile devices will account for 66% of the global mobile data traffic by 2014 as per forecasts [1]. Video services that have gained wide interest are so many and television broadcast, DVD, Blu-Ray, Mobile TV, Web TV etc. are some to name. One of the key characteristics of video services is the quality of experience (QoE) as observed by the end user. Quality of visual media can get degraded while capturing, storing, transmission, reproduction and display due to the distortions which might occur at any of these stages. Although many automatic (objective) methods of visual quality assessment (VQA) have been proposed and are in-use but the true judges of the quality are humans as end users. Subjective assessment of video quality is done by following standardized recommendations for experimental set-up, lab environment, stimuli characteristics and the number of viewers.

Video Quality Expert Group (VQEG), formed in 1997, has been the principal body for conducting comparative study of objective methods by performing subjective assessment tests. The full reference television (FR-TV) project report of phase I [2] concluded that no objective measure of VQA can replace subjective VQA. The test data and the corresponding subjective mean opinion score (MOS) was released to public for facilitating research on VQA. In the phase II campaign of subjective tests [3], it was found that some of the candidates of objective VQA performed better than PSNR, the conventional quality measure still in use. However, the MOS and test data was not made accessible for everyone, hence researchers in the field of development of objective VQA can not use the VQEG phase II data for the verification of their algorithms. Even the data which was shared earlier constitutes video sequences of high resolutions like 720x576 @50 fps and 720x486 @60 fps tested in PAL and NTSC TV formats respectively. Some independent efforts of subjective VQA have also been made such as [4], [5], [6] but unfortunately the test stimuli and subjective scores are not shared to public.

Tremendous efforts have been made to develop objective methods of VQA which are easy to repeat and not so time consuming as subjective assessment. Objective methods are categorized as full reference (FR), reduced reference (RR) and no reference (NR) methods based on the reference information required. Examples of each type include [7], [8], [9] for FR, RR and NR respectively. Most of the objective methods estimate visual quality by quantifying spatial (intra frame) degradations such as blocking, blurring, ringing and temporal (inter frame) degradations such as jitter [10]. High amount of quantization and frame dropping or resolution reduction to meet rate requirements of limited capacity transmission sources generate the aforementioned artifacts. However, the objective methods of VQA have to validated through subjective VQA for the appraisal of their performance. To study the effect of varying bitrate and spatio-temporal conditions on the video quality, only limited work has been performed and as mentioned before, the shared data lacks in variety of resolution. To the best of authors' knowledge, publically available databses of subjective VQA are only found from:

- [11], where the test stimuli for various coding conditions of H.264/AVC is limited.
- [12], that doesn't address H.264/AVC coded videos.
- [6], that has shared MOS values for H.264/AVC coded data but the stimuli was not made public.
- [13], that has shared MOS values for H.264/AVC coded data but the shared stimuli was impaired by transmission error only.

Moreover, in most of the reported work, either the coding details are not shared completely or the video encoding is done using bi-predictive coding which makes the decoding process slower. Real time applications like video-conferencing and popular video services such as videos viewed on handheld devices require low delay in the coding process. In this paper, we present a study of subjective VQA on QCIF (176x144)

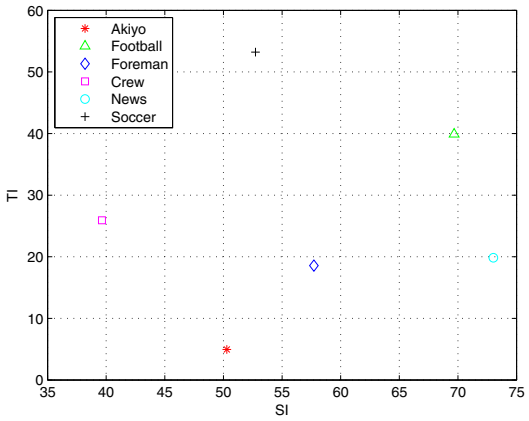


Fig. 1. SI and TI plot computed for luminance component of selected videos

and CIF (352x288) resolution videos encoded by H.264/AVC standard at various bitrate and frame rate conditions. To be applicable for low delay services, bi-predictive coding is not preferred. We believe that sharing this kind of data has fundamental role in facilitating comparison and benchmarking of objective methods of VQA. In the sequel, Section II presents details on the generation of test stimuli and the subjective assessment. Finally, Section III presents conclusive discussion on this work.

II. Subjective Assessment of the Test Stimuli

For a purposeful subjective VQA database, the test stimuli should constitute videos with varying amounts of spatial and temporal spectral information and the tests should be conducted by following standard recommendations. This section provides details of our approach on both of these considerations.

A. Generation of Test Stimuli

In order to characterize the spectral contents of a video sequence, spatial spectral information (SI) and temporal spectral information (TI) indices are used as suggested in the ITU T-recommendations [14]. SI and TI are calculated in the luminance plane of a video and the used formulae are given below.

$$SI = max_{time}[std_{space}[sobel(f_n)]] \quad (1)$$

- $sobel(f_n)$ is each frame at time n filtered with Sobel filter.
- std_{space} is standard deviation over pixels.
- max_{time} is maximum value in time series.

$$TI = max_{time}[std_{space}[M(i,j)] \quad (2)$$

Where $M(i,j)$ is difference of motion between pixel values in space for sequential frames in the luminance plane and (i,j) represent the pixel location. High textured videos have higher values of SI and higher motion content in a video sequence leads to higher TI value.Videos sequences selected for this work cover a wide range in both of these indices as depicted in Figure 1 namely *Akiyo, News, Foreman, Crew, Soccer, and Football* without any audio signal. One selected frame from each video is shown in Figure 2. These videos were in raw (YUV) progressive format with 4:2:0 color space having 300 frames in QCIF and CIF resolutions. Thus using these 12 source (Source Reference Circuit, the SRC) files, 120 test sequences (Hypothetical Reference Circuit, the HRC) were generated at ten different bitrates and two different frame rates with JM reference software for H.264/AVC, available online at [15]. In particular, the baseline profile was employed that is suitable for low delay applications. Table I presents details on the generated test sequences. For subjective VQA, all 120 HRCs were scaled to CIF size at frame rate of 30 fps. *Bicubic interpolation* and *repeat frame* methods were used for up-scaling QCIF videos to CIF and 15 fps to 30 fps respectively.

TABLE I
Description of Used Videos

Spatial Resolutions	Video Sequences	Bit rates(kbps)
CIF @ 30,15 frame rates	Akiyo	200,400,600,800,1000
	News	200,400,600,800,1000
	Foreman	200,400,600,800,1000
	Crew	200,400,600,800,1000
	Soccer	200,400,600,800,1000
	Football	200,400,600,800,1000
QCIF @ 30,15 frame rates	Akiyo	100,200,300,400,500
	News	100,200,300,400,500
	Foreman	100,200,300,400,500
	Crew	100,200,300,400,500
	Soccer	100,200,300,400,500
	Football	100,200,300,400,500

B. Subjective Assessment

We have followed the recommendations given by ITU-R BT 500-12 [16] for performing our experiments of subjective VQA. Particularly, the method followed was single stimulus quality evaluation where a test video sequence is shown once without presence of any explicit reference, corresponds to the reality where users see only the processed version of a video [17][18]. A flat LCD screen with non-glare surface treatment was used for displaying the video sequences. The used monitor had resolution 1440x900 with 5 ms response time and its color temperature was set at 6500K in sRGB mode. Other hardware includes a desktop HP system having 3 GHz AMD processor and 4 GB RAM. A comfortable seating arrangement was made for the subjects at a distance of three to four times the hight of the display size of a video. A software tool developed at the department was used to automate the process of presenting the videos in the center of the screen. Videos were played in a random order for each subject with insertion of the standard intervals (10 sec.) in between for grading. Viewers were not given the privilege to repeat any video and software front end had no controls available for the subjects to alter the intended processes in anyway. The used grading scale was 0-100 to have scores in a continuous manner. The software automatically stored the results in an excel sheet. In total, 21 non-expert

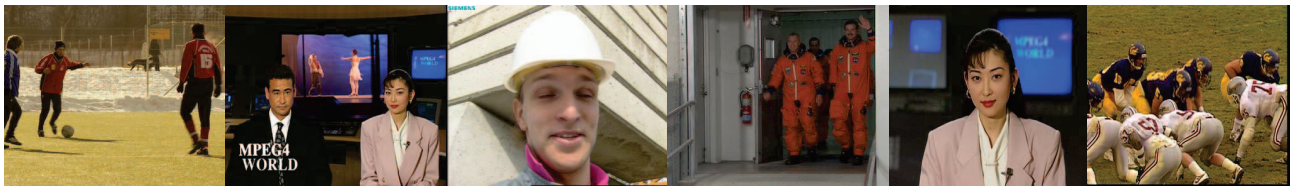


Fig. 2. A snapshot of the test video sequences

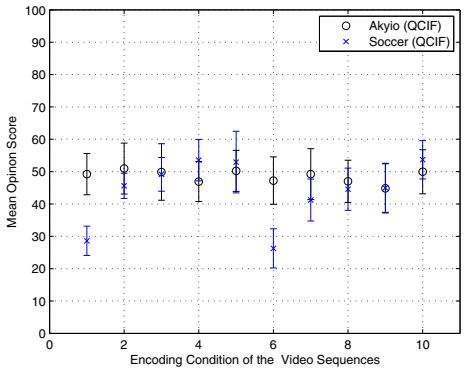
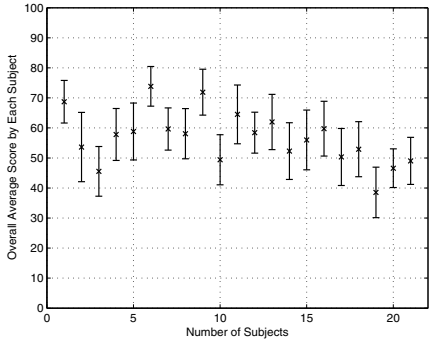
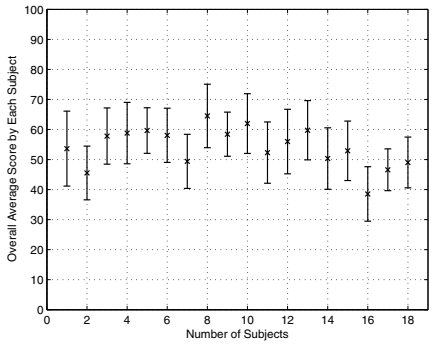


Fig. 3. MOS Values for Akiyo and Soccer



Raw Scores



Refined Scores

Fig. 4. Overall Average Score by Each Subject

subjects were invited to participate in the tests. The subjects were international students at different master programmes offered at the university and some staff members also took part in the grading campaign, both male and female. The viewers were introduced to the tests by dictating a common text saying that they are supposed to grade a set of videos on visual quality basis. To avoid any viewer fatigue, the test sessions were kept around half an hour length. Only one subject viewed the test stimuli at one time and the lab environment was kept silent to avoid any distractions for the subjects.

In order to obtain reliable results out of raw subjective scores, a two step filtering method was employed to refine the results. The first step was to detect and discard the scores from viewers who exhibited large change of votes compared to the average scores. The second step involved the screening of inconsistent observers without any thought of systematic change. The algorithmic details of these steps are reported in Annex 2 of [16]. After performing the refining process, the outliers were removed and we were left with scores by 18 subjects. Mean opinion score (MOS) was calculated from the scores of these subjects for each test condition k as following, where x represents score given by one subject and N is the number of subjects after outlier removal.

$$MOS_k = \frac{1}{N}\sum_{n=1}^{N} x_k(n) \tag{3}$$

As the number of the subjects in the refined scores is not large, we can assume that our data follows Students t distribution. Thus confidence interval (CI) estimates for each test condition is calculated using the following expression, where α is taken equal to 0.05 for having 95% CI, N is the number of subjects after outliers removal and σ_k is the standard deviation of a test condition k across N subjects.

$$CI_k = t(1-\alpha/2, N)\sigma_k \frac{1}{\sqrt{N}} \tag{4}$$

Figure 4 shows average score graded by each subject for all the videos before and after the refining process, within 95% CI. It is observed that variability of the scores among the subjects has been reduced by the refining process. Figure 3 presents the MOS values of Akiyo and Soccer sequences at 5 bitrate conditions for QCIF resolution videos as given in table I. First 5 values are obtained at frame rate of 30 fps and the next 5 values are obtained at frame rate of 15 fps. Overall, MOS values have an increasing trend with increase of bitrate

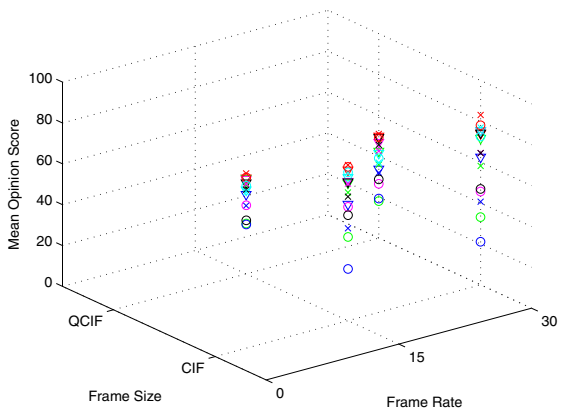


Fig. 5. MOS Values Plotted Against Frame Rate and Frame Size

TABLE II
POINTERS USED IN FIG. 5

Sequence/Coding Condition	1	2	3
Akiyo	▽	×	○
Crew	▽	×	○
Football	▽	×	○
Foreman	▽	×	○
News	▽	×	○
Soccer	▽	×	○

at both of the frame rates.

MOS values for coding condition 1: 600 kbps for CIF and 300 kbps for QCIF, coding condition 2: 400 kbps for CIF and 200 kbps for QCIF, coding condition 3: 200 kbps for CIF and 100 kbps for QCIF are plotted against frame rates and frame resolution in Figure 5. The plotted values conform to the commonly known tendencies of perceptual quality at varying bitrates and frame rates. For further illustration, MOS values for a set of test videos are plotted in Figure III for CIF and in Figure IV for QCIF resolutions. Table III and IV presents the complete list of MOS values for CIF and QCIF test videos.

III. CONCLUSION

This paper provides details on a subjective VQA of 120 video sequences coded by following H.264/AVC standard at different bitrates, frame rates and frame resolutions. Selection of the test stimuli was performed based on the values of spatial and temporal spectral information. The subjective tests were performed in accordance with standard recommendations of ITU-T. The MOS values and the compressed bistreams are shared online at: *http://www.bth.se/ing/shm.nsf/pages/research-resources*.
Our motive is to contribute to the field by sharing the complete test database with the fellow researchers. Related future works includes performing similar experiments on other frame resolutions and frame rates also. Moreover, the impact of other distortions can also be included such as packet loss and delay in the IP networks.

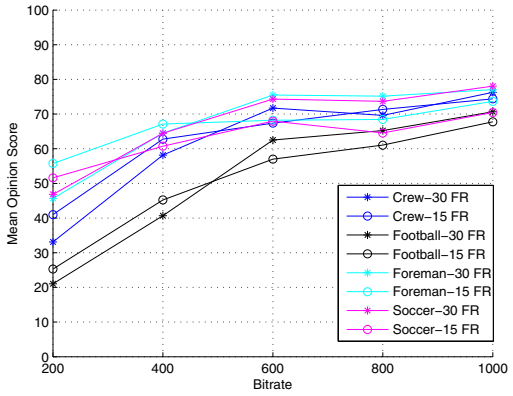


Fig. 6. MOS Values for Some Test CIF Videos

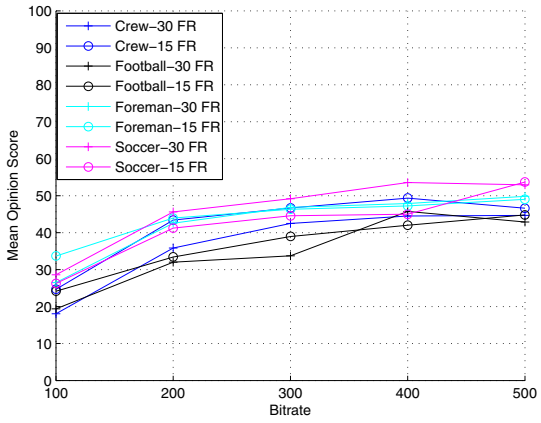


Fig. 7. MOS Values for Some Test QCIF Videos

REFERENCES

[1] "Visual networking index: Global mobile data traffic forecast update, 2009-2014," Feb 2010, cisco, Inc.

[2] "Final rep. from the video quality experts group on the validation of objective models of video quality assessment vqeg," 2000. [Online]. Available: www.vqeg.org

[3] "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii (fr-tv2)," 2003. [Online]. Available: www.vqeg.org

[4] S. Winkler and R. Campos, "Video quality evaluation for Internet streaming applications," in *Proc. IS&T/SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, vol. 5007, 2003, pp. 104–115.

[5] G.-M. Muntean, P. Perry, and L. Murphy, "Subjective assessment of the quality-oriented adaptive scheme," *Broadcasting, IEEE Transactions on*, vol. 51, no. 3, pp. 276 – 286, sept. 2005.

[6] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *Multimedia, IEEE Transactions on*, vol. 10, no. 7, pp. 1316 –1324, nov. 2008.

[7] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600 –612, april 2004.

[8] X. Wang, G. Jiang, and M. Yu, "Reduced reference image quality assessment based on contourlet domain and natural image statistics," in *Image and Graphics, 2009. ICIG '09. Fifth International Conference on*, sept. 2009, pp. 45 –50.

TABLE III
MOS VALUES FOR CIF RESOLUTION VIDEOS

Test Sequence	Frame rate	Bitrate	MOS
Akiyo	30	200	78.05
		400	83.11
		600	74.05
		800	76.50
		1000	76.55
	15	200	75.33
		400	76.22
		600	73.72
		800	73.50
		1000	68.55
Crew	30	200	33.11
		400	58.16
		600	71.72
		800	69.61
		1000	76.27
	15	200	41
		400	62.77
		600	67.38
		800	71.33
		1000	74.38
Football	30	200	21.05
		400	40.66
		600	62.5
		800	65.22
		1000	70.66
	15	200	25.27
		400	45.22
		600	57
		800	61.05
		1000	67.77
Foreman	30	200	45.66
		400	64.50
		600	75.5
		800	75.16
		1000	77.05
	15	200	55.72
		400	67.11
		600	68.22
		800	68.44
		1000	73.66
News	30	200	72.55
		400	76.94
		600	75.61
		800	75.72
		1000	80
	15	200	72.72
		400	72.72
		600	70.05
		800	73.50
		1000	71.83
Soccer	30	200	46.83
		400	64.4
		600	74.33
		800	73.66
		1000	78.05
	15	200	51.61
		400	60.72
		600	67.94
		800	64.55
		1000	70.44

TABLE IV
MOS VALUES FOR QCIF RESOLUTION VIDEOS

Test Sequence	Frame rate	Bitrate	MOS
Akiyo	30	100	49.22
		200	50.94
		300	49.88
		400	46.94
		500	50.22
	15	100	47.22
		200	49.22
		300	47
		400	44.83
		500	49.94
Crew	30	100	18.05
		200	35.83
		300	42.50
		400	44.5
		500	44.66
	15	100	24.66
		200	43.38
		300	46.66
		400	49.33
		500	46.61
Football	30	100	19.38
		200	32
		300	33.72
		400	45.72
		500	42.88
	15	100	24.16
		200	33.38
		300	38.94
		400	42
		500	44.77
Foreman	30	100	26.50
		200	42.55
		300	46.83
		400	47.88
		500	49.83
	15	100	33.66
		200	43.94
		300	46.33
		400	47.22
		500	49
News	30	100	39.11
		200	36.94
		300	41.44
		400	42.11
		500	43.16
	15	100	43.11
		200	40.05
		300	41.11
		400	42.16
		500	42.94
Soccer	30	100	28.61
		200	45.61
		300	49.16
		400	53.55
		500	52.94
	15	100	26.27
		200	41.22
		300	44.55
		400	45
		500	53.66

[9] M. Shahid, A. Rossholm, and B. Lövström, "A reduced complexity no-reference artificial neural network based video quality predictor," in *Image and Signal Processing (CISP), 2011 4th International Congress on*, vol. 1, oct. 2011, pp. 517 –521.

[10] M. Yuen and H. R. Wu, "A survey of hybrid mc/dpcm/dct video coding distortions," *Signal Process.*, vol. 70, no. 3, pp. 247–278, Nov. 1998.

[11] L. C. H.R. Sheikh, Z.Wang and A. Bovik, "Live image quality assessment database release 2." [Online]. Available: http://live.ece.utexas.edu/research/quality

[12] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, "Color image database for evaluation of image quality metrics," in *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, oct. 2008, pp. 403 –408.

[13] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A h.264/avc video database for the evaluation of quality metrics," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, march 2010, pp. 2430 –2433.

[14] "Subjective video quality assessment methods for multimedia applications," September 1999, iTU-T, Recommendation ITU-R P910.

[15] JVT, "H.264/MPEG-4 AVC Reference Software," *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6), 24th Meeting: Geneva, CH, 29*, 29 June - 5 July 2007. [Online]. Available: http://iphome.hhi.de/suehring/tml/download

[16] "ITU-R Radio communication Sector of ITU, Recommendation ITU-R BT.500-12," 2009.

[17] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. Wiley, 2005.

[18] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *Broadcasting, IEEE Transactions on*, vol. 57, no. 1, pp. 1 –14, march 2011.