A H.264/AVC VIDEO DATABASE FOR THE EVALUATION OF QUALITY METRICS*

F. De Simone[a], M. Tagliasacchi[b], M. Naccari[b], S. Tubaro[b], T. Ebrahimi[a]

[a] Ecole Politechnique Fédérale de Lausanne, Multimedia Signal Processing Group,
CH-1015 Lausanne, Switzerland
[b] Politecnico di Milano, Dipartimento di Elettronica e Informazione,
20133 Milano, Italy

ABSTRACT

This paper describes a publicly available database of subjective scores, relative to quality assessment of 156 video streams encoded with H.264/AVC and corrupted by simulating packet losses over an error-prone network. The data has been collected in controlled test environments at the premises of two academic institutions. A detailed statistical analysis of subjective results has been performed, showing high consistency of the collected scores. In addition to subjective scores, we have made available to the research community both the uncompressed files and the H.264/AVC bitstreams of each video sequence, in order to provide a common database of benchmark data to test and compare the performance of full-reference, reduced-reference and no-reference video quality assessment algorithms.

Index Terms— Subjective test, video quality assessment, H.264/AVC, packet losses.

1. INTRODUCTION

Research in the field of video quality assessment relies on the availability of subjective scores, collected by means of experiments where groups of people are asked to rate the quality of video sequences. In order to gather reliable and statistically significant data, subjective tests have to be carefully designed and performed, and require a relevant number of subjects. For these reasons, the subjective tests are usually very time consuming. Nevertheless, subjective data are fundamental to test and compare the performance of the objective algorithms, i.e. metrics, which try to predict human perception of video quality by automatically analyzing the video streams.

Examples of comparative studies of objective video quality metrics are those carried out by VQEG (Video Quality Expert Group) [1] and [2], based on the results of two extensive campaigns of subjective tests which involved many laboratories. Unfortunately, the subjective results and the test material used to perform these studies have not been made publicly available, thus VQEG subjective results cannot be used by independent researchers for testing of more recent and future metrics. Also, many studies are available in literature, reporting results of subjective experiments, such as [3] and [4] which investigate quality degradation in video streaming applications, but none of them has provided public access to the collected data to the research community.

At the best of author's knowledge, the only publicly available databases of subjective results and related test material, in the field of visual quality assessment, are the LIVE database [5] and the TID2008 database [6] for standard definition images, and the EPFL database for high resolution images [7]. The first contribution of a public database for video quality evaluation has been proposed by the authors in [8]. This database includes CIF video sequences coded with H.264/AVC reference software and impaired by simulating packet losses over an error-prone channel. The subjective data have been gathered at the premises of two academic institutions: Politecnico di Milano (PoliMI) - Italy, and Ecole Politechnique Fédérale de Lausanne (EPFL) - Switzerland. In addition to subjective data, the database includes the original video contents and configuration files used to encode them, the original and corrupted H.264/AVC bitstreams, as well as the network simulator used to generate the test material.

Our goal in this paper is to present an extension of the study in [8], by including also video contents at 4CIF resolution and to discuss the statistical analysis of the subjective data gathered by the two aforementioned institutions. The rest of the paper is organized as follows: Section 2 briefly describes the test material, the environmental setup and the subjective evaluation methodology used in our study; Section 3 presents the statistical analysis performed on the subjective data in order to describe the results of the experiments and to compare the data of the two laboratories; Section 4 concludes the paper.

2. SUBJECTIVE TEST CAMPAIGN

In order to produce meaningful subjective results, the test material needs to be carefully selected and the subjective evaluation procedure must be rigorously defined. In our work, we adapted the specifications given in [9] and [10] as detailed in [8] and summarized in the following.

2.1. Test video sequences

To produce the test material, we considered twelve video sequences in raw progressive format. Six sequences, namely *Foreman*, *Hall*, *Mobile*, *Mother*, *News* and *Paris*, have CIF spatial resolution (352×288 pixels) and frame rate equal to 30 fps. The other six sequences, namely *Ice*, *Harbour*, *Soccer*, *CrowdRun*, *DucksTakeoff* and *ParkJoy*, have 4CIF spatial resolution (704×576 pixels): the former three sequences are available at 30 fps; the latter three sequences were obtained by cropping HD resolution video sequences down to 4CIF resolution and downsampling the original content from 50 fps to 25 fps. These sequences were selected since they are representative of different levels of spatial and temporal complexity, as computed by means of the Spatial Information (SI) and Temporal
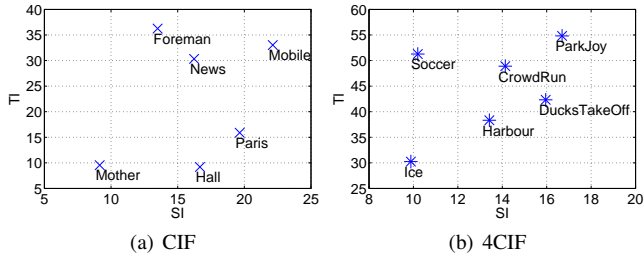
(a) CIF (b) 4CIF

Fig. 1. Spatial Information (SI) and Temporal Information (TI) indexes computed on the luminance component of the selected video sequences [11].

Table 1. Quantization Paramaters used to encode the video sequences and corresponding bitrates and PSNR values.

Sequence	Spatial res. and fps	Bitrate [kbps]	PSNR [dB]	QP
Foreman	CIF 30fps	353	34.4	32
News	CIF 30fps	283	37.3	31
Mobile	CIF 30fps	532	28.3	36
Mother	CIF 30fps	150	37.0	32
Hall monitor	CIF 30fps	216	36.2	32
Paris	CIF 30fps	480	33.6	32
Ice	4CIF 30fps	1325	40.8	28
Soccer	4CIF 30fps	2871	37.2	28
Harbour	4CIF 30fps	5453	36.3	28
CrowdRun	4CIF 25fps	6757	33.47	30
DucksTakeOff	4CIF 25fps	7851	30.43	34
ParkJoy	4CIF 25fps	6187	31.46	32

Information (TI) indexes [11] and showed in Figure 1.

We encoded all sequences using the H.264/AVC [12] High Profile to enable B-pictures and Context Adaptive Binary Arithmetic Coding (CABAC) for coding efficiency. Each frame is divided into a fixed number of slices, where each slice consists of a full row of macroblocks. The rate control available in the reference software was disabled since it introduced visible quality fluctuations along time for some of the video sequences. Instead, a fixed Quantization Parameter (QP) was carefully selected for each sequence so as to ensure high visual quality in absence of packet losses. The QP values used for each sequence, along with the corresponding bitrates and PSNR values, are shown in Table 1. Apart from considering the PSNR values as indicators of the quality level, each coded sequence was visually inspected in order to check whether the chosen QPs minimized the blocking artifacts induced by lossy coding. For further details of the coding configuration please refer to [8].

For each of the twelve original H.264/AVC bitstreams, we generated a number of corrupted bitstreams by dropping packets according to a given error pattern. To simulate burst errors, the patterns have been generated at six different PLRs [0.1%, 0.4%, 1%, 3%, 5%, 10%] with a two state Gilbert's model with an average burst length of 3 packets, since it is characteristic of IP networks [13]. For each PLR and content, two realizations were selected, thus a total of 72 CIF test sequences and 72 4CIF test sequences were included in the test material.

2.2. Subjective evaluation procedure

The CIF and 4CIF data were presented in two separate test sessions. Subjects were seated directly in line with the center of the video display at a specified viewing distance, equal to 6-8H for CIF data and 4-6H for 4CIF data, where H is the native height of the video frame. Laboratories set up and environmental conditions are detailed in [8].

In our subjective evaluation we adopted a Single Stimulus (SS) method in which a processed video sequence is presented alone, without being paired with its unprocessed reference version. The test procedure included also the reference version of each video sequence, which in this case is the packet loss free sequence, as a free-standing stimulus for rating like any other.

Each sequence was displayed for 10 seconds. At the end of each test presentation, follows a 3-5 seconds voting time, when the subject rates the quality of the stimulus using the 5 point ITU continuous scale in the range $[0-5]$ as described in [11].

Each session included 83 video sequences: 72 test sequences, i.e. realizations corresponding to 6 different contents and 6 different PLRs; 6 reference sequences, i.e. packet loss free video sequences; 5 stabilizing sequences, i.e. dummy presentations, shown at the beginning of the test session to stabilize observers' opinion. The dummy presentations consist in 5 realizations, corresponding to 5 different quality levels, selected from the same test sequences. The results for these items are not registered by the evaluation software but the subject is not told about this.

The presentation order for each subject is randomized according to a random number generator, discarding those permutations where stimuli related to the same original content are consecutive.

Before each test session, written instructions are provided to the subjects to explain their task. Additionally, a training session is performed to allow the viewer to familiarize with the assessment procedure and the software user interface. The contents shown in the training session are not used in the test session and the data gathered during the training are not included in the final test results. In particular, for the training phase we used two different contents for each spatial resolution, namely *Coastguard* and *Container* at CIF resolution, *City* and *Crew* at 4CIF resolution, and 5 realizations of each, representatives of the 5 point ITU continuous scale.

The number of subjects involved in the test sessions are as follows: twenty-three for CIF and twenty-one for 4CIF at PoliMI and seventeen for CIF and nineteen for 4CIF at EPFL. All subjects reported that they had normal or corrected to normal vision. Their age ranged from 24 to 40 years old. Some of the subjects were Ph.D. students working in fields related to image and video processing, some were naive subjects.

3. STATISTICAL ANALYSIS OF THE RESULTS

The statistical analysis of the results aims at studying whether: i) the variation in subjective scores is a result of the intended variation of experimental variables, rather than a random variation; ii) the conclusions of a study based on a limited sample of subjects are valid for the entire population. The scores collected for the CIF and 4CIF sessions by the two laboratories have been processed separately but applying the same procedure. Prior to any processing, we used the Shapiro-Wilk test to verify the normality of distributions of scores across subjects. The results of this test showed that the scores distributions are normal or close to normal, thus justifying the processing detailed in the next subsections.

3.1. Data screening

Although each subject has been trained according to the same procedure, subjects may have used the rating scale differently. This behaviour can be modeled by representing the raw score m_{ij} assigned by the subject i for the test condition j as:

$$m_{ij} = g_i m_j + o_i + n_{ij} \qquad (1)$$

where m_j is the true quality score for the stimulus j, g_i is a gain factor, o_i is an offset, and n_{ij} is a sample from a zero-mean, white Gaussian noise [14]. In this model, the gain and offset could vary from subject to subject. If the variations are large across the subjects or the number of subjects is small, a normalization procedure can be used to reduce the gain and the offset variations among test subjects.

In order to check the between subject variability, a two-way ANOVA was applied to the raw scores [14]. Under the null hypothesis for between subjects variation, scores given by the various subjects are samples drawn from the same distribution. The p-value resulting from the ANOVA expresses the probability of observing the obtained scores if the null hypothesis was true. Therefore, a sufficiently small p-value suggests that the null hypothesis can be firmly rejected. The p-value for between subjects variation computed on the raw scores was always equal to zero, showing that there were significant differences among mean scores of different subjects. Thus, a subject-to-subject correction was applied, according to the following rule [14]:

$$m'_{ij} = \left(m_{ij} - \bar{m}_i + \mu \right) / \left(4\sigma_i / K \right) \qquad (2)$$

with the score after normalization m'_{ij}, the mean $\bar{m}_i$ and the standard deviation σ_i computed for each subject i across the test conditions, the overall mean μ across all subjects and test conditions, and K a scaling factor equal to the upper limit value of the rating scale. An example of the effect of the normalization over score distribution is shown in Figure 2, where the boxplot of the raw scores obtained at EPFL for CIF data before and after normalization are shown.

After the normalization, outliers have been detected according to the guidelines described in section 2.3.1 of Annex 2 of [10].

3.2. Mean opinion scores and confidence intervals

After screening, the results of the test campaign can be summarized by computing the mean opinion score (MOS) for each test condition j (i.e. combination of video content and PLR) as:

$$MOS_j = \frac{\sum_{i=1}^{N} m'_{ij}}{N} \qquad (3)$$

where N is the number of subjects after outliers removal and m'_{ij} is the score by subject i for the test condition j after normalization.

The relationship between the estimated mean values based on a sample of the population (i.e. the subjects who took part in our experiments) and the true mean values of the entire population is given by the confidence interval of estimated mean. Due to the small number of subjects, the $100 \times (1 - \alpha)\%$ confidence intervals (CI) for mean opinion scores were computed using the Students t-distribution, as follows:

$$CI_j = t(1 - \alpha/2, N) \cdot \frac{\sigma_j}{\sqrt{N}} \qquad (4)$$

where $t(1 - \alpha/2, N)$ is the t-value corresponding to a two-tailed t-Student distribution with $N - 1$ degrees of freedom and a desired significance level α (equal to 1-degree of confidence). N corresponds to the number of subjects after outliers detection, and σ_j
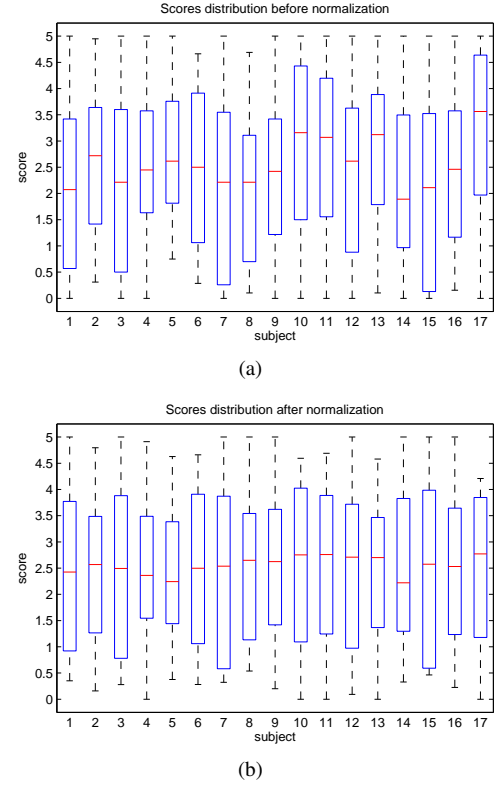

(a)

(b)

Fig. 2. Effect of the normalization over EPFL scores for CIF data.

is the standard deviation of a single test condition across subjects. The interpretation of a confidence interval is that if the same test is repeated for a large number of times, using each time a random sample of the population, and a confidence interval is constructed every time, then $100 \times (1 - \alpha)\%$ of these intervals will contain the true value. We computed our confidence intervals for an α equal to 0.05, which corresponds to a degree of significance of 95%.

An example of resulting plots showing PoliMI MOS values and associated confidence intervals for one CIF and one 4CIF content, namely *News* and *DucksTakeoff*, is presented in Figure 3. Here, the horizontal axis indicates the PLRs together with the considered realization ("a" or "b"). The plots for the entire set of contents considered in the study are available for download [1]. As a general comment, the MOS plots resulting from the study, like that in Figure 3, clearly show that the experiments have been properly designed, since the subjective rates uniformly span the entire range of quality levels. Also, the confidence intervals are reasonably small, thus, prove that the effort required from each subject was appropriate and subjects were consistent in their choices.

3.3. Comparing results of the two laboratories

The most straightforward way to compare the results obtained in the two independent laboratories for each video content is to show obtained MOS and CI values in the same plot, along with the corresponding scatter plot of MOSs, as shown in Figure 4 for *Foreman* content. The scatter plot and the correlation coefficients give indication of the excellent degree of correlation among the results of the

---

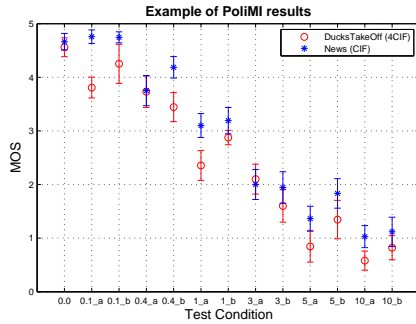[1] http://mmspg.epfl.ch/vqa or http://vqa.como.polimi.it

Fig. 3. MOS values and 95% confidence interval obtained by PoliMI laboratory for *Foreman* CIF content and *Ice* 4CIF content.

two laboratories. Particularly the Pearson's coefficient measures the scatter of the points around the linear trend, while the Spearman's coefficient measures the monoticity of the mapping, i.e. how well an arbitrary monotonic function describes the relationship among the two set of data. The results for the other video sequences are available for download.
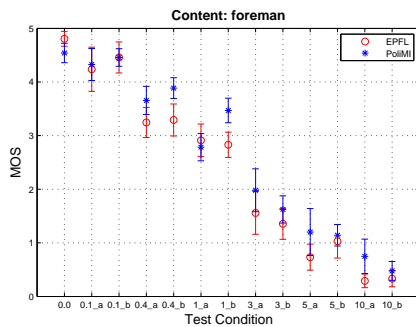
(a)
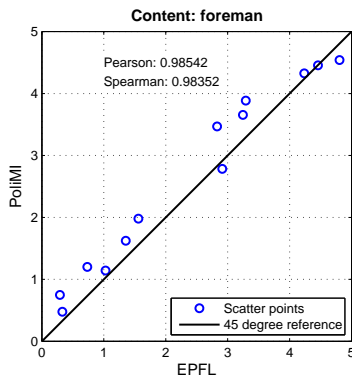
(b)

Fig. 4. (a) MOS values and 95% confidence interval obtained by PoliMI and EPFL laboratories for *Foreman* CIF content and (b) related scatter plot and correlation coefficients.

4. CONCLUSION

In this paper the procedure followed in order to produce a publicly available dataset of subjective results for 156 CIF and 4CIF video sequences has been described. The test material, the tools used to produce it, and the subjective results of the study are available for public download at `http://mmspl.epfl.ch/vqa` and `http://vqa.como.polimi.it`. The results of the subjective tests performed in two different laboratories show high consistency and correlation.

We believe that such a publicly available database will allow easier comparison and performance evaluation of the existing and future objective metrics for quality evaluation of video sequences, contributing to the advance of the research in the field of objective quality assessment.

5. REFERENCES

[1] "Final report from the video quality experts group on the validation of objective models of video quality assessment," Tech. Rep., Video Quality Expert Group, March 2000.

[2] P. Corriveau and A. Webster, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," Tech. Rep., Video Quality Expert Group, July 2003.

[3] S. Winkler and R. Campos, "Video quality evaluation for Internet streaming applications," in *Human Vision and Electronic Imaging*, Santa Clara, CA, USA, January 2003.

[4] G.-M. Muntean, P. Perry, and L. Murphy, "Subjective assessment of the quality-oriented adaptive scheme," *IEEE Trans. Broadcast.*, vol. 51, no. 3, pp. 276–286, September 2005.

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

[6] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, and F. Battisti, "Color image database for evaluation of image quality metrics," in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, Cairns, AU, October 2008.

[7] F. De Simone, L. Goldmann, V. Baroncini, and T. Ebrahimi, "Subjective evaluation of JPEG XR image compression," in *SPIE, Vol. 7443*, San Diego, CA, USA, August 2009.

[8] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," in *First International Workshop on Quality of Multimedia Experience*, San Diego, CA, USA, July 2009.

[9] "VQEG Hybrid Testplan, Version 1.2," ftp://vqeg.its.bldrdoc.gov.

[10] ITU-T, *Recommendation ITU-R BT 500-10*, March 2000, Methodology for the subjective assessment of the quality of the television pictures.

[11] ITU-T, *Recommendation ITU-R P 910*, September 1999, Subjective video quality assessment methods for multimedia applications.

[12] "H.264/AVC reference software version JM14.2," Tech. Rep., Joint Video Team (JVT), downloadable at http://iphome.hhi.de/suehring/tml/download/.

[13] T.-K. Chua and D. C. Pheanis, "QoS evaluation of sender-based loss-recovery techniques for VoIP," *IEEE Netw.*, vol. 20, no. 6, pp. 14–22, December 2006.

[14] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, Iowa State University Press, 1989.