Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2011, Article ID 190431, 12 pages
doi:10.1155/2011/190431

Research Article

Subjective Quality Assessment of H.264/AVC Video Streaming with Packet Losses

Francesca De Simone,[1] Matteo Naccari,[2] Marco Tagliasacchi,[3] Frederic Dufaux,[4] Stefano Tubaro,[3] and Touradj Ebrahimi (EURASIP Member)[1]

[1] *Multimedia Signal Processing Group (MMSPG), Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland*
[2] *Instituto de Telecomunicações, Instituto Superior Técnico, 1049-011 Lisboa, Portugal*
[3] *Dipartimento di Elettronica e Informazione, Politecnico di Milano (PoliMI), 20133 Milano, Italy*
[4] *Telecom ParisTech, 75634 Paris Cedex 13, France*

Correspondence should be addressed to Francesca De Simone, francesca.desimone@epfl.ch

Received 15 November 2010; Accepted 18 January 2011

Academic Editor: Vittorio Baroncini

Research in the field of video quality assessment relies on the availability of subjective scores, collected by means of experiments in which groups of people are asked to rate the quality of video sequences. The availability of subjective scores is fundamental to enable validation and comparative benchmarking of the objective algorithms that try to predict human perception of video quality by automatically analyzing the video sequences, in a way to support reproducible and reliable research results. In this paper, a publicly available database of subjective quality scores and corrupted video sequences is described. The scores refer to 156 sequences at CIF and 4CIF spatial resolutions, encoded with H.264/AVC and corrupted by simulating the transmission over an error-prone network. The subjective evaluation has been performed by 40 subjects at the premises of two academic institutions, in standard-compliant controlled environments. In order to support reproducible research in the field of full-reference, reduced-reference, and no-reference video quality assessment algorithms, both the uncompressed files and the H.264/AVC bitstreams, as well as the packet loss patterns, have been made available to the research community.

1. Introduction

The use of IP networks for video delivery is gaining an increasing popularity as a mean of broadcasting data from content providers to consumers. Video transmission over peer-to-peer networks is also becoming very popular. Typically, these networks provide only best-effort services, that is, there is no guarantee that the content will be delivered without errors. In practice, the received video sequence may be a degraded version of the original. Besides distortions introduced by lossy coding, user's experience might be also affected by channel-induced distortions. Thus, the design of systems for automatic monitoring of the received video quality is of great interest for service providers, in order to optimize the transmission strategies as well as to ensure a desired level of quality of experience. Several algorithms, usually referred to as objective quality metrics, have been proposed in the literature for the in-service objective quality evaluation of video sequences. They consist of No-Reference or Reduced-Reference methods relying on the analysis of the bitstream, the pixels, or both (so-called hybrid approach) [1, 2]. Nevertheless, the lack of publicly available databases of video sequences and subjective scores makes the comparison of existing and novel solutions very difficult.

In fact, research in the field of video quality assessment relies on the availability of subjective scores, collected by means of experiments in which 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 large number of subjects. For these reasons, the subjective tests are usually very time consuming. Nevertheless, the availability of subjective scores is fundamental to enable validation and comparative benchmarking of objective video

quality metrics in a way to support reproducible and reliable research results.

The first public database of video contents and related subjective quality scores was produced by the Video Quality Experts Group (VQEG) and used to compare the performance of Full-Reference objective metrics, targeting secondary distribution of television as application [3]. Unfortunately, only part of the subjective results and test materials used to perform the study has been made publicly available. Additionally, this dataset includes interlaced video sequences and focuses on MPEG-2 compression. These distortions are not representative of the current video coding and transmission technologies. Thus, the usage of VQEG data by independent researchers to validate more recent and future metrics is limited. Recently, two video databases have been proposed by the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin. The LIVE Video Quality Database [4] consists of a set of video sequences corresponding to different contents, distorted by MPEG-2 and H.264/AVC compression as well as by transmission over error-prone IP and wireless networks. The presence of diverse distortion types makes this database particularly useful to test the consistency of metrics performance. The LIVE Wireless Video Quality Assessment Database [5] focuses on distortions due to transmission over a wireless network and takes into account a set of video sequences having similar content concerning airplanes. These databases include the test video sequences and the processed subjective results and have been used to evaluate the performance of a set of Full-Reference video quality metrics in [4, 5].

In this paper, a detailed description of the publicly available database originally presented in [6] and extended in [7] is provided, along with an extensive discussion of the data processing applied to the collected subjective scores and analysis of the results. The database focuses on the impact of packet losses on visual quality. It contains subjective scores collected through subjective tests carried out at the premises of two academic institutions: Ecole Polytechnique Fédérale de Lausanne-Switzerland and Politecnico di Milano-Italy. The same experiments were performed at both laboratories and a total of 40 subjects were asked to rate 144 video sequences, corresponding to 12 different video contents at CIF and 4CIF spatial resolutions and different Packet Loss Rates (PLRs), ranging from 0.1% to 10%. The packet loss free sequences were also included in the test material, thus in total 156 sequences were rated by each subject at each institution.

With respect to others cited above, the database described in this paper, (1) includes data collected at the premises of two different laboratories, showing high correlation among the two sets of collected results, as an indicator of reliability of the subjective data as well as of the adopted evaluation methodology, (2) includes the decoded sequences and the compressed video streams affected by packet losses, as well as the packet loss patterns, thus it can be used for testing stream-based and hybrid No-Reference and Reduced-Reference metrics, (3) includes the complete set of collected subjective results, including the raw scores before any data processing, thus allowing reproducible research on subjective data processing and detailed statistical analysis of metrics performance. The database is available for download at http://mmspg.epfl.ch/vqa and http://vqa.como.polimi.it.

The rest of the paper is organized as follows. Section 2 describes the test material, the environmental setup, and the subjective evaluation methodology used in our study. In Section 3, the processing of the results is detailed. The results of the two laboratories are analyzed and compared in Section 4. Finally, Section 5 concludes the paper.

2. Subjective Video Quality Assessment

In a subjective video quality test, a group of people is asked to watch a set of video sequences and to rate their quality. The design of formal subjective experiments involves four main phases [8–10]:

(1) Selection of the Test Material. The test material has to be a realistic sample of the actual data that belongs to the target application scenario. Also, in order to avoid decreasing subject's level of attention, the content has to be heterogeneous and the test sessions should not last more than 30 minutes each, including any training phase. For the same reason, it is important to select stimuli whose quality levels are possibly uniformly distributed across the rating scale. Therefore, an accurate supervised screening of the test material is needed. Whenever it is not possible to show the entire set of the test materials in a single test session (for instance, because the duration of the session exceeds 30 minutes), multiple sessions may be scheduled, so that each subject is able to perform all the sessions and rate all the test material (i.e., full factorial design). Alternatively, a reduced subset of test conditions may be selected.

(2) Selection of the Test Methodology. Several internationally accepted test methods for subjective video quality assessment are described in [11, 12]. A first taxonomy of the methods regards how the visual stimuli are presented to the viewer. In Double Stimulus methods, the observer is sequentially presented with two video sequences: one of the two sequences is the reference stimulus and the other is the test stimulus. The observer can be asked to rate either both stimuli, or only the test stimulus. In Single Stimulus methods, only one stimulus is shown and has to be rated. Finally, in Stimulus Comparison methods, pairs of stimuli are shown simultaneously and the subject is asked to compare their quality. A second classification of test methodologies concerns the rating scale in which the subject is asked to express her/his quality evaluation score. A first distinction is between continuous and discrete scales. A second distinction pertains the use of either a categorical scale (textual labels, describing the quality of the stimulus or the annoyance of the impairments) or a numerical scale. Finally, the subject may be asked to enter her/his rating after the visualization of the test material, and/or directly while watching the video sequence. Such continuous evaluations can be used to elicit an indication of the temporal quality variations across the sequence.

(3) Selection of the Participants. In order to gather statistically significant data, subjective tests require a large enough number of subjects, as a representative sample of the population of interest. The participants to the test have to be screened for visual acuity and color blindness. They can be chosen from two categories of end users depending on the goal of the investigation: expert or naive, that is, nonexpert.

(4) Choice of the Experimental Setup. The experimental setup should reproduce the viewing conditions of the target application scenario, while keeping under control all the external experimental parameters which could influence subject's perception. Some recommendations and setup parameters are indicated in [11, 13]. An accurate control of the test environment is necessary to ensure the reproducibility of the test activity and to compare results across different laboratories and test sessions.In the following, the test material, the environment setup, the subjective evaluation methodology, and the panel of subjects used in our study are described.

2.1. Test Material. To produce the test material for the subjective evaluation campaign, twelve video sequences in raw progressive format and 4 : 2 : 0 chrominance subsampling ratio were considered. Six sequences, namely, *Foreman*, *Hall*, *Mobile*, *Mother*, *News,* and *Paris*, had 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*, had 4CIF spatial resolution (704 × 576 pixels). The former three sequences were 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 because they represented different levels of spatial and temporal complexity. The complexity was quantified by means of Spatial Information (SI) and Temporal Information (TI) indexes [12]. The SI and TI indexes computed on the luminance component of each sequence [12] are shown in Figure 1. The first frame of each test sequence is shown in Figures 2 and 3. Furthermore, four additional sequences, two for each spatial resolution, were used for training, as detailed in Section 2.3, namely, *Coastguard* and *Container* at CIF resolutions, *City* and *Crew* at 4CIF resolutions. All sequences were 10 seconds long.

Before simulating packet losses, the sequences were compressed using the H.264/AVC reference software, version JM14.2, available for download at [14]. All sequences were encoded using the High Profile to enable B-pictures and Context Adaptive Binary Arithmetic Coding (CABAC) for coding efficiency. Each frame was divided into a fixed number of slices, where each slice consisted of a full row of macroblocks. The rate control was disabled, as 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. Each coded sequence was visually inspected in order to check whether the chosen QPs minimized the blocking artifacts induced by lossy coding. Table 1 illustrates the parameters used to generate the compressed bitstreams and Table 2 the bit-rates and PSNR values corresponding to the selected QPs for all the test sequences.

Table 1: H.264/AVC encoding parameters.

Reference software	JM14.2
Profile	High
Number of frames	298
Chroma format	4 : 2 : 0
GOP size	16
GOP structure	*IBBPBBPBBPBBPBB*
Number of reference frames	5
Slice mode	Fixed number of macroblocks
Rate control	Disabled, fixed QP (Table 2)
Macroblock partitioning for motion estimation	Enabled
Motion estimation algorithm	Enhanced Predictive Zonal Search (EPZS)
Early skip detection	Enabled
Selective intramode decision	Enabled

For each of the twelve original H.264/AVC bitstreams, a number of corrupted bitstreams were generated, by dropping packets according to a given error pattern [15]. Coded slices belonging to the first frames were not corrupted, as they contained header information (Picture Parameter Set (PPS) and Sequence Parameter Set (SPS)). Conversely, the remaining slices might be discarded from the coded bitstream. To simulate burst errors, the patterns were generated at six different PLRs, 0.1%, 0.4%, 1%, 3%, 5%, 10%, with a two-state Gilbert's model [16]. The model parameters were tuned to obtain an average burst length of 3 packets, which is a typical characteristic of IP networks [17]. The two-state Gilbert's model generated, for each PLR, several error patterns. For each PLR and content, two decoded video sequences were manually selected in order to uniformly span a wide range of distortions, that is, perceived video quality, while keeping the size of the dataset manageable. The details of the selection procedure can be found in [6]. A total of 72 CIF sequences with packet losses and 72 4CIF sequences with packet losses were included in the test material. Each bitstream was decoded with the H.264/AVC reference software decoder with motion-compensated error concealment turned on [18].

2.2. Environment Setup. Each test session involved only one subject per display assessing the test material. The CIF and 4CIF sequences 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 6–8 H for CIF sequences and to 4–6 H for 4CIF sequences [13], where H denotes the native height of the video window in the screen. Table 3 summarizes the specifications of the display devices. The ambient lighting system in both laboratories consisted of neon lamps with color temperature of 6500 K.

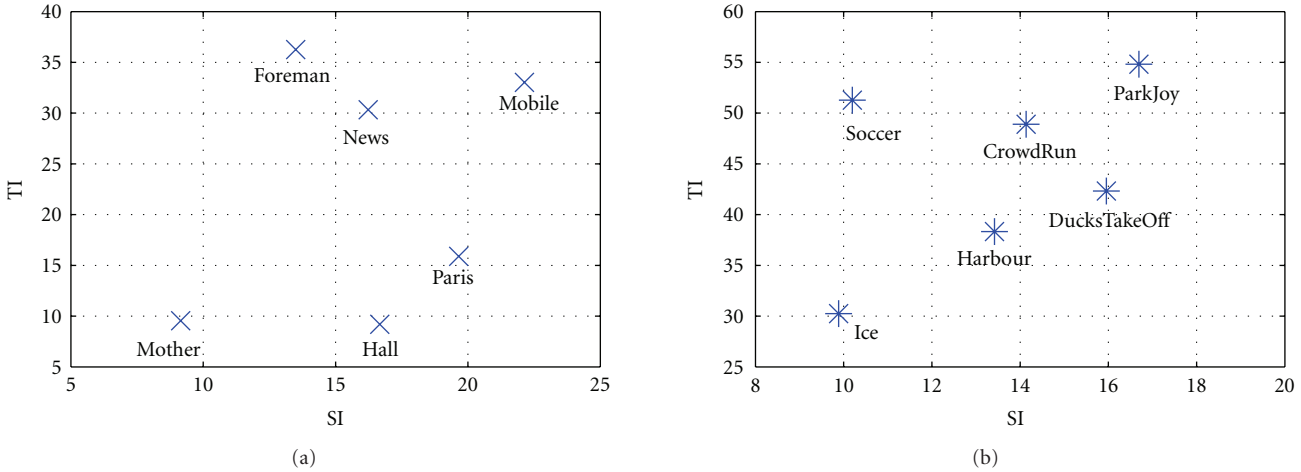


Figure 1: Spatial Information (SI) and Temporal Information (TI) indexes computed on the luminance component of the selected (a) CIF and (b) 4CIF video sequences [12].

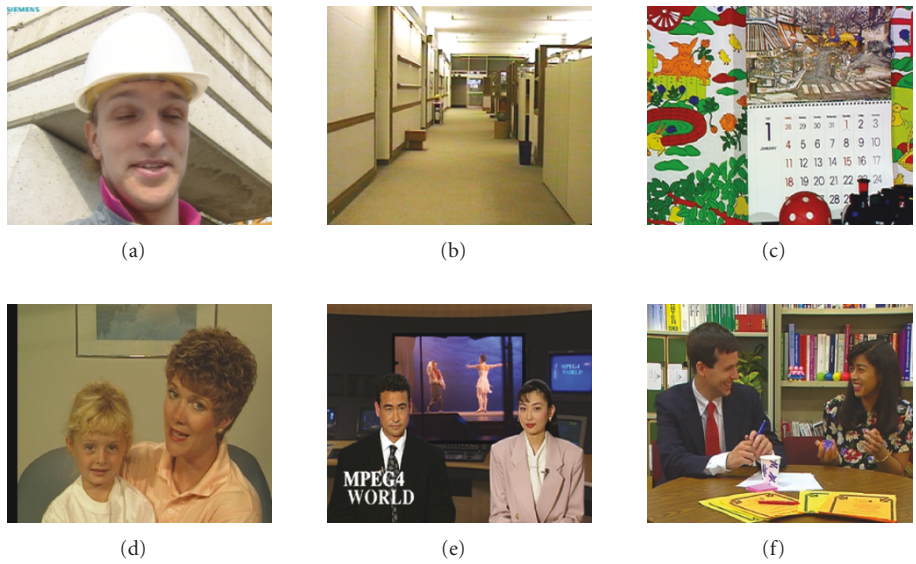
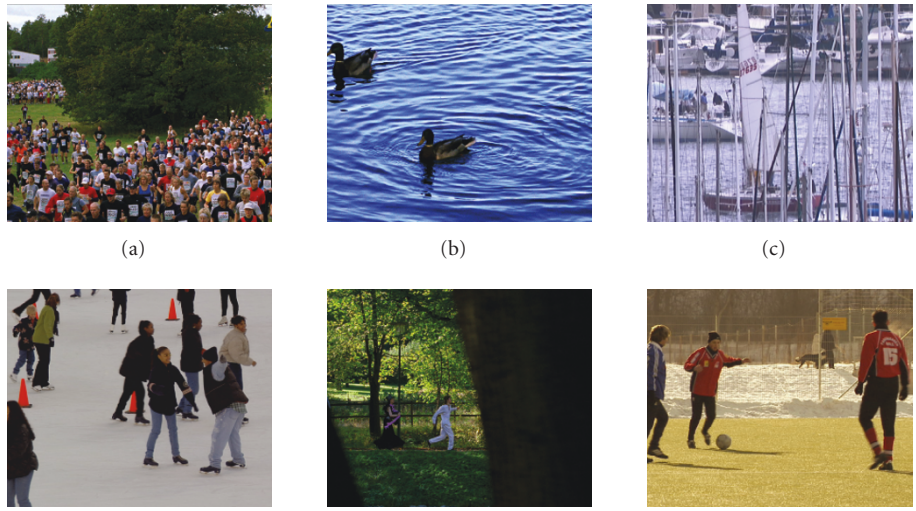


Figure 2: First frame of each CIF test sequence: (a) Foreman, (b) Hall, (c) Mobile, (d) Mother, (e) News, and (f) Paris.

Table 2: Test sequences and coding conditions.

Sequence name	Spatial res. and fps	MB/slice	Bit-rate (kbps)	PSNR (db)	QP
Foreman	CIF 30 fps	22	353	34.4	32
News	CIF 30 fps	22	283	37.3	31
Mobile	CIF 30 fps	22	532	28.3	36
Mother	CIF 30 fps	22	150	37.0	32
Hall	CIF 30 fps	22	216	36.2	32
Paris	CIF 30 fps	22	480	33.6	32
Ice	4CIF 30 fps	44	1325	40.8	28
Soccer	4CIF 30 fps	44	2871	37.2	28
Harbour	4CIF 30 fps	44	5453	36.3	28
CrowdRun	4CIF 25 fps	44	6757	33.4	30
DucksTakeOff	4CIF 25 fps	44	7851	30.4	34
ParkJoy	4CIF 25 fps	44	6187	31.4	32



(a) (b) (c)

(d) (e) (f)

Figure 3: First frame of each 4CIF test sequence: (a) Crowdrun, (b) DucksTakeOff, (c) Harbour, (d) Ice, (e) Parkjoy, and (f) Soccer.

Table 3: Specifications of LCD display devices.

	EPFL	PoliMI
Type	Eizo CG301W	Samsung SyncMaster 920N
Diagonal size	30 inches	19 inches
Resolution	2560×1600 (native)	1280×1024 (native)
Calibration tool	EyeOne Display 2	EyeOne Display 2
Gamut	sRGB	sRGB
White point	D65	D65
Brightness	120 cd/m^2	120 cd/m^2
Black level	minimum	minimum

2.3. Test Methodology. The Single Stimulus (SS) method was used to collect the subjective data. Thus, each processed video sequence was presented alone, without being paired with its unprocessed, that is, reference, version. However, the test procedure included a reference version of each video sequence, which in this case was the packet loss free sequence, as a freestanding stimulus for rating like any other. At the end of each test presentation, a voting time followed, when the subjects were asked to rate the quality of the stimulus using a five-point ITU continuous adjectival scale.

A dedicated Matlab-based GUI was developed to present the stimuli and the rating scale. The video sequences were shown at their native resolutions, centered in a grey-128 background window at full screen. To show the uncompressed video sequences without adding temporal impairments due to rendering latency, an optimized media player [19] was used, embedded into the GUI, and the workstations of the two laboratories were equipped with high performance video servers. Figure 4 shows the rating window presented to the subjects. It was decided not to limit the voting time and to present the next stimulus only after pressing the "Done" button. The subject could not proceed

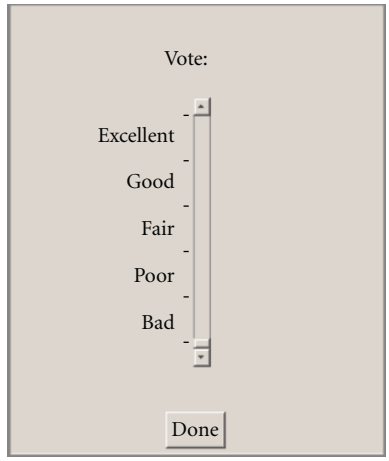


Figure 4: Five-point continuous adjectival quality scale [12].

with the test unless she/he entered a score. Note that although a continuous adjectival quality scale in the range 0 to 5 was adopted, the numerical values were used only for data analysis and were not shown to the subjects.

Each test session referred to a single spatial resolution (i.e., either CIF or 4CIF) and included 83 video sequences: 6×12 test sequences, that is, realizations corresponding to 6 different contents and 6 different PLRs; 6 reference sequences, that is, packet loss free video sequences; 5 stabilizing sequences, that is, dummy presentations shown at the beginning of the experiment to stabilize observers' opinion. The dummy presentations consisted in 5 realizations, corresponding to 5 different quality levels, selected from the test video sequences. The results for these items were not registered by the evaluation software but the subject was not told about this. The presentation order for each subject was randomized, discarding those permutations where stimuli related to the same original content were consecutive.

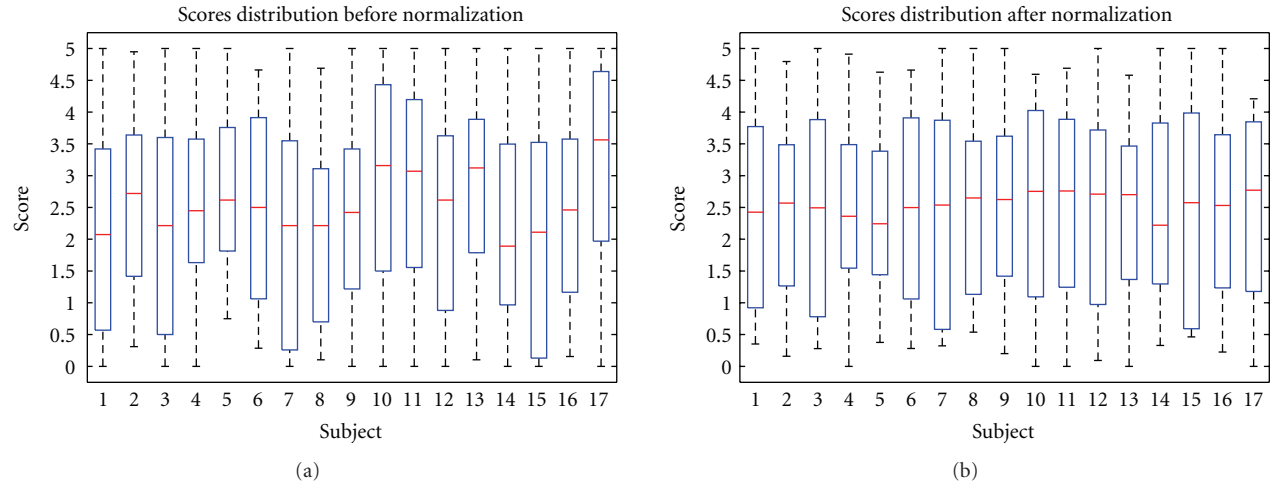


Figure 5: Effect of the normalization over EPFL scores for CIF data: distribution of the raw data (a) before and (b) after normalization.

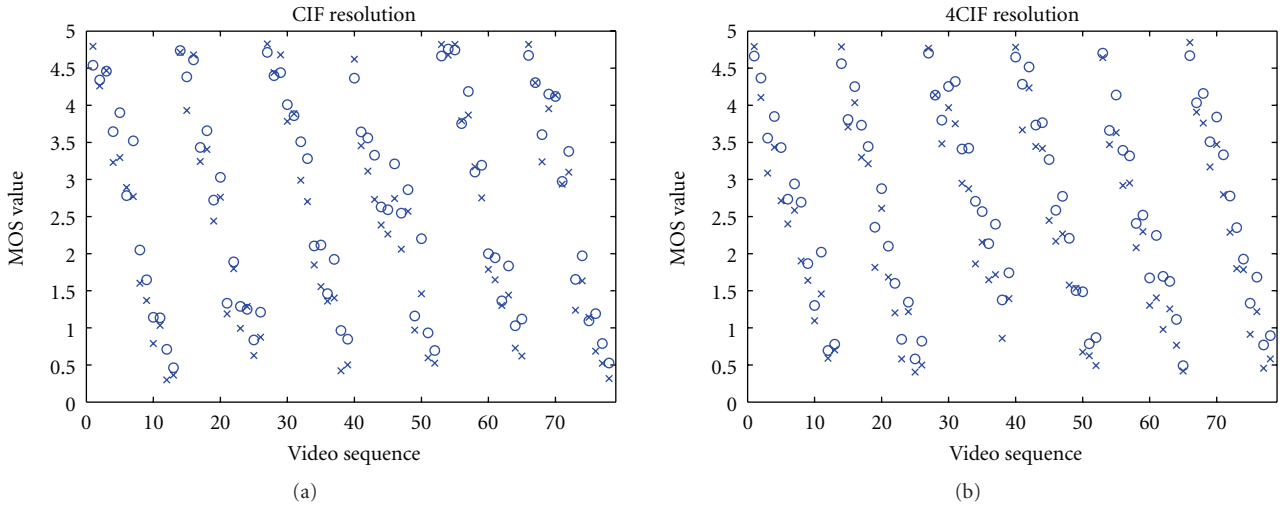


Figure 6: Distribution of MOS values obtained by PoliMI (o) and EPFL (x) laboratories for (a) CIF content and (b) 4CIF content.

Before each test session, written instructions were provided to subjects to explain their task. Additionally, a training session was performed to allow viewers to familiarize with the assessment procedure and the software user interface. The contents shown in the training session were not used in the test sessions and the data gathered during the training was not included in the final test results. In particular, for the training phase two different contents for each spatial resolution and five realizations of each, representatives of the score labels depicted in Figure 4, were used. During the display of each training sequence, the operator explained the meaning of each label, as summarized in the written instructions reported in Appendix A.

The entire set of test sessions was performed in each laboratory. Twenty-three and twenty-one subjects took part in the CIF and 4CIF sessions, respectively, at PoliMI. Seventeen and nineteen subjects took part in the CIF and 4CIF sessions, respectively, at EPFL. All subjects reported that they had normal or corrected to normal vision. Their age ranged from 24 to 40 years old. All observers were naive subjects.

3. Subjective Data Processing

Some general guidelines for processing the results of quality assessment experiments are detailed in [11] but, when dealing with subjective data, the statistical tools to be used need to be selected according to the properties of the data under analysis. Thus, a case-by-case approach is preferable.

In general, the results of a subjective experiment for quality assessment are summarized by averaging the scores assigned by the panel of observers to each video sequence, that is, stimulus, in order to obtain a Mean Opinion Score (MOS) and corresponding Confidence Interval (CI) [11]. Prior to the MOS computation, a correction procedure, that is, normalization, to compensate for any systematic differences in the usage of the rating scale by the different

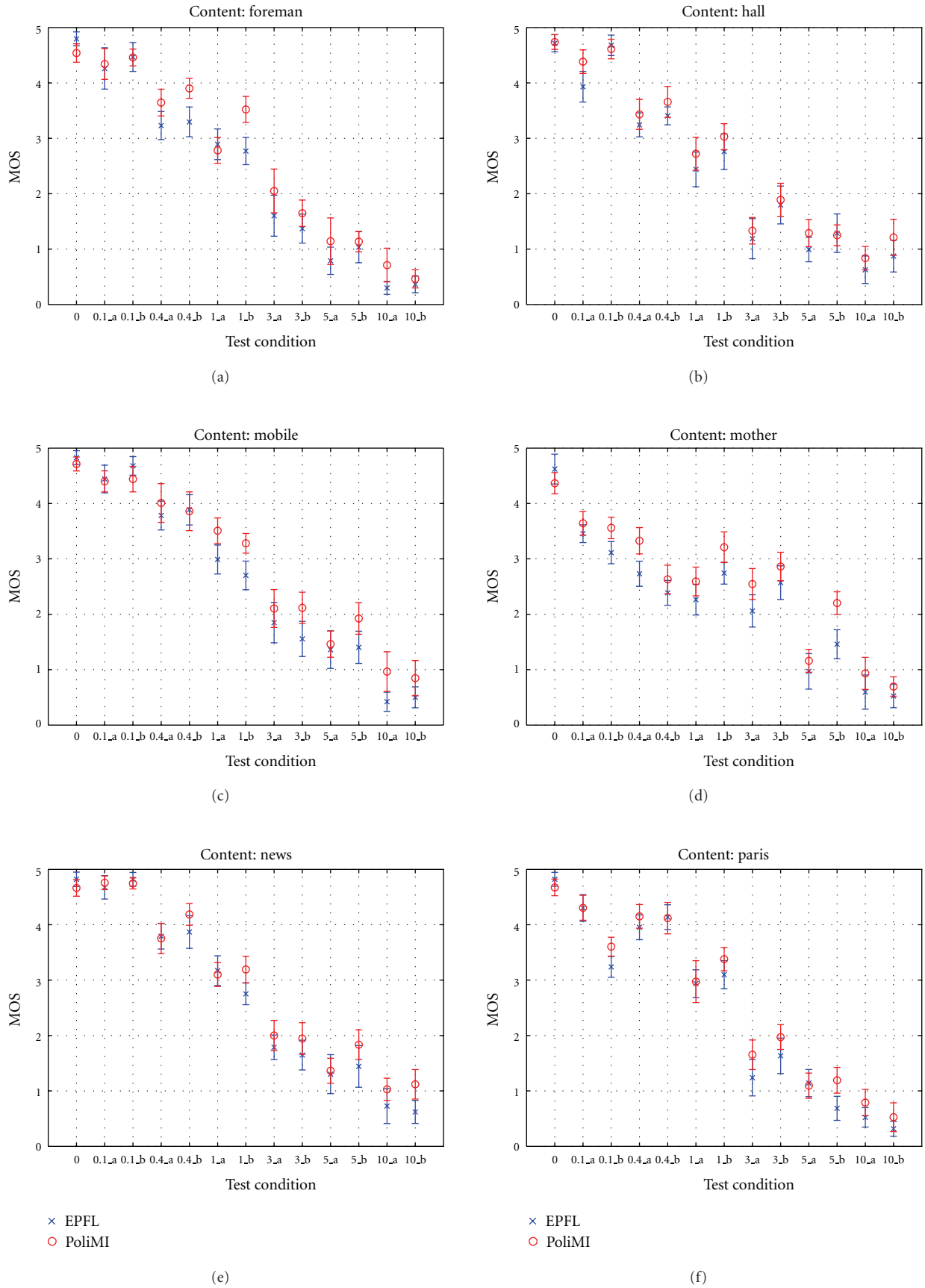


Figure 7: MOS values and 95% confidence intervals obtained by PoliMI and EPFL laboratories for CIF contents.

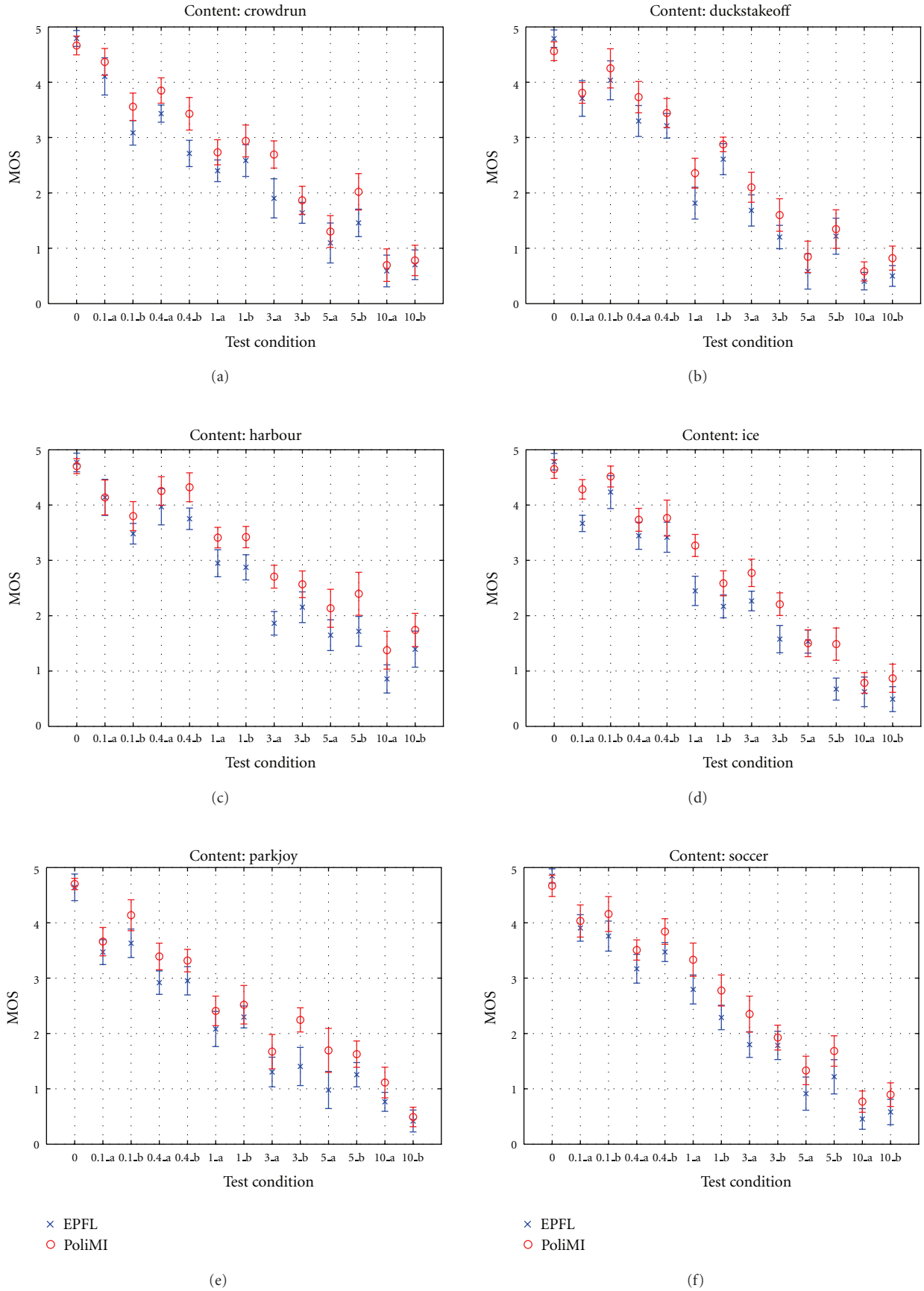


Figure 8: MOS values and 95% confidence intervals obtained by PoliMI and EPFL laboratories for 4CIF contents.

subjects can be applied. Then, the scores are screened in order to detect and exclude possible outliers, that is, subjects whose scoring significantly deviates from others.

The scores collected in the CIF and 4CIF sessions by the two laboratories were processed separately, according to the procedure detailed below.

3.1. Scores Normalization. First, in order to check the between-subject variability, a two-way ANOVA was applied to the raw scores [20]. The results of the ANOVA showed that a subject-to-subject correction was needed. Thus the scores were normalized according to offset mean correction rule. The details of the normalization procedure are described in Appendix B. This approach differs from that applied in [4, 5, 21], where the normalization procedure consists in converting the scores to z-scores, mainly because not only the mean score of the single subject is taken into account to correct his/her own scores, but also the overall mean across all test conditions and subjects.

An example of the effect of the normalization over the scores distribution is shown in Figure 5, where the boxplot of the raw scores obtained at EPFL for CIF data before and after normalization are shown.

3.2. Subjects Screening. The screening of possible outlier subjects was performed considering the normalized scores, according to the procedure described in Appendix C. Three and one outliers were detected out of 23 and 17 subjects, from the results produced for CIF data at PoliMI and at EPFL, respectively. Four and two outliers were detected out of 21 and 19 subjects, from the results produced for 4CIF data at PoliMI and at EPFL, respectively. The scores corresponding to the outlier subjects were discarded from the results.

3.3. Mean Opinion Scores and Confidence Intervals. After the screening, the results of the test campaign were summarized by computing the MOS for each test condition j (i.e., combination of video content and PLR):

$$\mathrm{MOS}_j = \frac{1}{N}\sum_{s=1}^{N} m_{sj} \tag{1}$$

with N the total number of subjects after outlier removal and m_{sj} the score assigned by subject s to the test condition j, after normalization. Finally, 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 was computed as the confidence interval of estimated mean. Because of the small number of subjects, the 95% confidence intervals (δ) for the mean subjective scores were computed using the Student's t-distribution, as follows:

$$\delta = t_{(1-\alpha/2)} \cdot \frac{S}{\sqrt{N}}, \tag{2}$$

where $t_{(1-\alpha/2)}$ is the t-value associated with the desired significance level α for a two-tailed test ($\alpha = 0.05$) with $N-1$ degrees of freedom, where N denotes the number of observations in the sample (i.e., the number of subjects after outliers detection) and S the estimated standard deviation of the sample of observations.

4. Analysis of the Results

The MOS values obtained for the entire set of video sequences by two laboratories are shown in Figure 6. These plots clearly show that the experiments have been properly designed, as the subjective rates uniformly span over the entire range of quality levels. Figures 7 and 8 show, for each video content, the MOS and CI values obtained after the processing applied to the subjective scores collected at PoliMI and at EPFL. The confidence intervals are reasonably small, thus, showing that the effort required from each subject was appropriate and subjects were consistent in their choices.

Additionally, as it can be noticed from the plots, there exists a good correlation between the data collected by the two laboratories. The most straightforward way to compare the results obtained in the two independent laboratories is to analyze the scatter plot of MOS values shown in Figure 9 for the entire set of video sequences. The scatter plot and the correlation coefficients give an indication of the excellent degree of correlation between the results of the two laboratories. The Pearson coefficient measures the distribution of the points around the linear trend, while the Spearman coefficient measures the monotonicity of the mapping, that is, how well an arbitrary monotonic function describes the relationship between two sets of data. The scatter plots show that the data from PoliMI are usually slightly shifted towards higher estimated quality levels, when compared to the results obtained at EPFL. The same trend can be observed in the raw scores, thus it can be concluded that it is an intrinsic property of the scores.

Additionally a Hotellings T^2-test for two series of population means [22] has been applied, separately to the results of the CIF and the 4CIF experiments, to understand whether the data of the two laboratories could be merged to compute overall MOS and associated CI values to be used as benchmark values for example for testing the performance of objective metrics. The null hypothesis is of no difference between the two multivariate patterns of scores, that is the subjects in the two laboratories do not differ in their responses to the stimuli. The result of the test for both the resolutions indicates that the the null hypothesis cannot be rejected, as a further proof of the fact that the two datasets of results could be merged for future studies involving the subjective results.

Finally, it is worth mentioning that the experiment for the quality assessment of CIF sequences was also carried out by performing the same evaluation under uncontrolled environmental conditions, in order to analyze the effect of the environment on the subjective scoring. A laptop was used to show the GUI and the test room was a normal office or living room. Twelve subjects took part in the experiments. Unfortunately, the results do not allow drawing any conclusion upon a systematic effect of the environmental conditions on the scores, since according to the content under analysis a different behavior of the MOS values, with

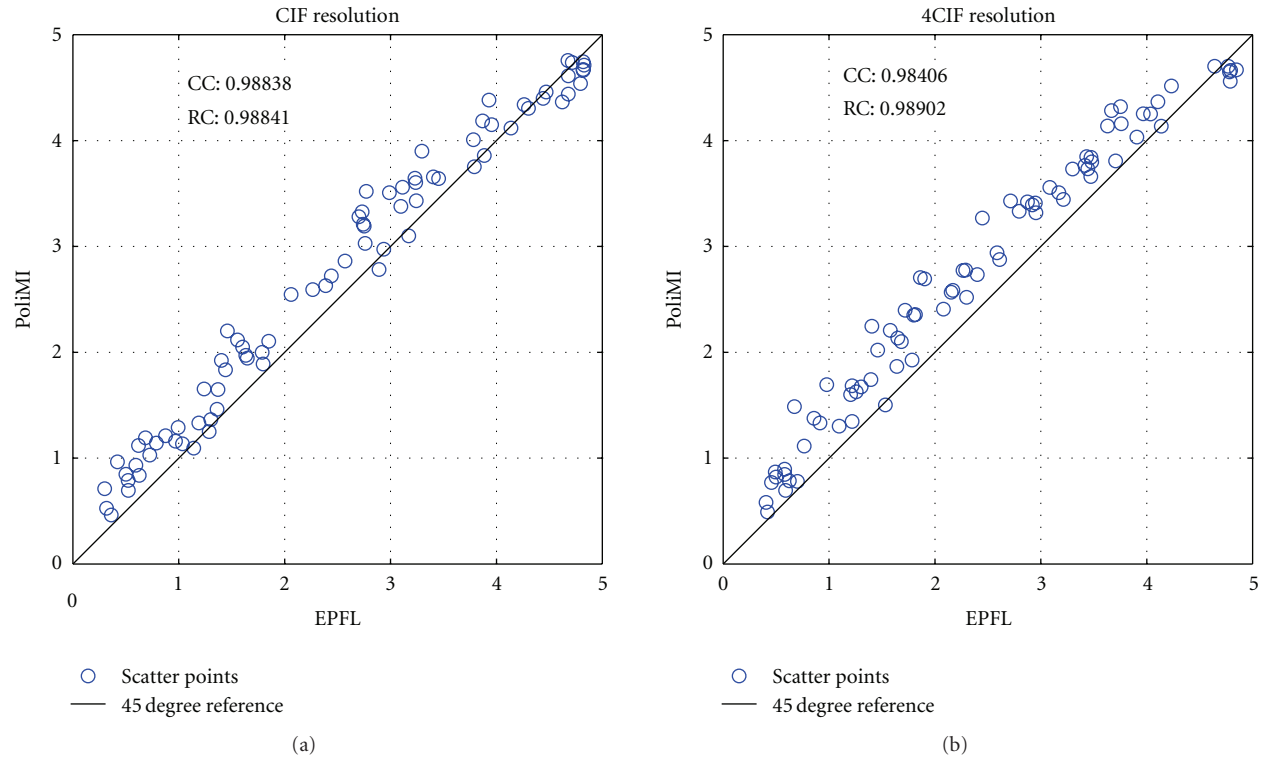


Figure 9: Scatter plots and Pearson (CC) and Spearman (RC) correlation coefficients between the MOS values obtained at PoliMI and EPFL for (a) CIF content and (b) 4CIF content.

respect to the results of the formal evaluations, was noticed. Currently an investigation is in progress in order to better understand the mechanisms behind the variability of the obtained results.

5. Concluding Remarks

In this paper, a database, containing the results of a subjective evaluation campaign aiming at studying the subjective quality of video sequences affected by packet losses, is presented. Subjective data was collected at the premises of two institutions, Ecole Polytechnique Fédérale de Lausanne and Politecnico di Milano. The database is publicly available at http://mmspg.epfl.ch/vqa and http://vqa.como.polimi.it and contains data relative to 156 sequences, both at CIF and 4CIF spatial resolutions. More specifically, the following data are provided: (1) the test material, together with the software tools used to produce them, (2) the corresponding H.264/AVC bitstreams, useful for evaluating no-reference and reduced-reference metrics, (3) the raw subjective scores, (4) the final MOS and CI data, together with the Matlab implementation of the algorithms adopted for score normalization and outlier screening.

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.

Future works will focus on an extension of the current database in order to include distortions due to jitter and delay. Also, further investigation on the effect of the environmental conditions over the results of the subjective experiments will be performed, focusing on the mobile scenario, where the assessment of video quality for test material at CIF spatial resolution is more realistic.

Appendices

A. Training Instructions

> "*In this experiment you will see short video sequences on the screen that is in front of you. Each time a sequence is shown, you should judge its quality and choose one point on the continuous quality scale.*"

(i) *Excellent*: the content in the video sequence may appear a bit blurred but no other artifacts are noticeable (i.e., only the lossy coding is present).

(ii) *Good*: at least one noticeable artifact is detected in the entire sequence.

(iii) *Fair*: several noticeable artifacts are detected, spread all over the sequence.

(iv) *Poor*: many noticeable artifacts and strong artifacts (i.e., artifacts which destroy the scene structure or create new patterns) are detected.

(v) *Bad*: very strong artifacts (i.e., artifacts which destroy the scene structure or create new patterns) are detected in the major part of the sequence.

B. Scores Normalization

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

$$m_{sc} = g_s m_c + o_s + n_{sc}, \tag{B.1}$$

where m_c is the true quality score for the stimulus c, g_s is a gain factor, o_s is an offset, and n_{sc} is a sample from a zero-mean, white Gaussian noise [20]. 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 [20]. Under the null hypothesis for between-subjects variation, scores given by 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-subject variation computed on the raw scores was always equal to zero, showing that there were significant differences between mean scores of different subjects. Thus, a subject-to-subject correction was applied, according to the following rule [20]:

$$m'_{sc} = K\left(\frac{m_{sc} - \overline{m}_s + \mu}{4S_s}\right), \tag{B.2}$$

with the score after normalization m'_{sc}, the mean $\overline{m}_s$ and the standard deviation S_s computed for each subject s 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.

C. Outlier Rejection

The goal of outlier rejection is to detect inconsistent subjects that show a significant bias of votes compared to the average behavior. The procedure, recommended in [11], can be summarized in the following steps:

(1) compute the kurtosis β index based on the scores assigned by each subject;

(2) if $2 \leq \beta \leq 4$ (i.e., score distribution is approximately normal) set $\omega = 2$ else set $\omega = \sqrt{20}$;

(3) for each content c do

(a) compute the mean score $\overline{m}'_c$ and the standard deviation S_c;

(b) set P_s equal to the number of times the score of one subject if above $\overline{m}'_c + \omega S_c$;

(c) set Q_s equal to the number of times the score of one subject if below $\overline{m}'_c - \omega S_c$;

(4) compute $a = (P_s + Q_s)/n$ and $b = |(P_s - Q_s)/(P_s + Q_s)|$

(5) if $a > 0.05$ and $b < 0.3$, then the subject is rejected.

Acknowledgments

This work has been partially sponsored by the EU under PetaMedia Network of Excellence and by the Swiss National Foundation for Scientific Research in the framework of NCCR Interactive Multimodal Information Management (IM2). Part of the material presented in this paper has been published in [6, 7].

References

[1] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: applications and human-motivated design," *Signal Processing*, vol. 25, no. 7, pp. 469–481, 2010.

[2] S. Winkler and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, 2008.

[3] 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, 2003.

[4] 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. 6, pp. 1427–1441, 2010.

[5] 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. 587–599, 2010.

[6] 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 *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEx '09)*, pp. 204–209, 2009.

[7] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '10)*, pp. 2430–2433, Dallas, Tex, USA, 2010.

[8] H. R. Wu and K. R. Rao, *Digital Video Image Quality and Perceptual Coding (Signal Processing and Communications)*, CRC Press, 2005.

[9] S. Winkler, *Digital Video Quality—Vision Model and Metrics*, John Wiley & Sons, New York, NY, USA, 2005.

[10] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method, and Application*, John Wiley & Sons, New York, NY, USA, 2006.

[11] ITU-T, "Recommendation ITU-R BT 500-10," Methodology for the subjective assessment of the quality of the television pictures, 2000.
[12] ITU-T, "Recommendation ITU-R P 910," Subjective video quality assessment methods for multimedia applications, 1999.
[13] VQEG hybrid testplan, version 1.2, ftp://vqeg.its.bldrdoc.gov.
[14] Joint Video Team (JVT), "H.264/AVC reference software version JM14.2," http://iphome.hhi.de/suehring/tml/download/.
[15] M. Luttrell, S. Wenger, and M. Gallant, "New versions of packet loss environment and pseudomux tools," Tech. Rep., Joint Video Team (JVT), 1999.
[16] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell System Technical Journal*, vol. 39, pp. 1253–1266, 1960.
[17] T. K. Chua and D. C. Pheanis, "QoS evaluation of sender-based loss-recovery techniques for VoIP," *IEEE Network*, vol. 20, no. 6, pp. 14–22, 2006.
[18] G. J. Sullivan, T. Wiegand, and K.-P. Lim, "Joint model reference encoding methods and decoding concealment methods," Tech. Rep. JVT-I049, Joint Video Team (JVT), 2003.
[19] Mplayer, http://www.mplayerhq.hu/.
[20] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, Iowa State University Press, 1989.
[21] 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, 2006.
[22] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, 1979.