SUBJECTIVE ASSESSMENT OF H.264/AVC VIDEO SEQUENCES TRANSMITTED OVER A NOISY CHANNEL*

F. De Simone [a], *M. Naccari* [b], *M. Tagliasacchi* [b], *F. Dufaux* [a], *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

In this paper we describe a database containing subjective assessment scores relative to 78 video streams encoded with H.264/AVC and corrupted by simulating the transmission over error-prone network. The data has been collected from 40 subjects at the premises of two academic institutions. Our goal is to provide a balanced and comprehensive database to enable reproducible research results in the field of video quality assessment. In order to support research works on Full-Reference, Reduced-Reference and No-Reference video quality assessment algorithms, both the uncompressed files and the H.264/AVC bitstreams of each video sequence have been made publicly available for the research community, together with the subjective results of the performed evaluations.

***Index Terms*—** Subjective video quality assessment, packet loss rate, H.264/AVC, error resilience.

1. INTRODUCTION

The use of IP networks for the delivery of multimedia contents is gaining an increasing popularity as a mean of broadcasting media files from a content provider to many content consumers. In the case of video, for instance, packet-switched networks are used to distribute programs in IPTV applications. Typically, these kinds of networks provide only best-effort services, i.e. there is no guarantee that the content will be delivered without errors to the final users. In some circumstances, the content provider and the user might decide to stipulate a Service Level Agreement (SLA) that fixes an expected perceived quality at the end-user terminal: the provider fixes a price to the customers for assuring the agreed Quality of Service (QoS), and pays a penalty if the SLA is unfulfilled. For this reason, it is fundamental in IP networks in particular, and video broadcasting applications in general, to assess the visual quality of distributed video contents.

In practice, the received video sequences may be a degraded versions of the original ones. Besides the distortion introduced by lossy coding, the user's experience might be affected by channel induced distortions. In fact, the channel might drop packets, thus introducing errors that propagate along the decoded video content because of the predictive nature of conventional video coding schemes [1, 2, 3], or it might cause jitter delay, due to decoder buffer underflows determined by network latencies.

With this contribution, we aim at providing a publicly available database containing Mean Opinion Scores (MOSs) collected during subjective tests carried out at the premises of 2 academic institutions: Politecnico di Milano - Italy and Ecole Polytechnique Fédérale de Lausanne - Switzerland. Fourty subjects were asked to rate 72 video sequences corresponding to 6 different video contents at CIF spatial resolution and different packet loss rates (PLR), ranging from 0.1% to 10%. The packet loss free sequences were also included in the test material, thus finally 78 sequences were rated by each subject. In this paper we address only the effect of packet losses and we refer the reader to the available literature [4][5] for aspects related to the effect of delay.

We emphasize that the availability of MOSs is fundamental to enable validation and comparative benchmarking of objective video quality assessment systems in such a way to support reproducible research results.

The rest of this paper is organized as follows. Section 2 introduces subjective quality assessment and illustrates the test material, environmental setup and subjective evaluation process used in our tests. Section 3 presents the processing over subjective data in order to normalize the collected scores and prune them from outliers. Section 4 presents the results and the correlation between the collected data in the two institutions and, finally, Section 5 concludes the paper.

2. SUBJECTIVE VIDEO QUALITY ASSESSMENT

In subjective tests, a group of subjects is asked to watch a set of video clips and to rate their quality. The scores assigned by the observers are averaged in order to obtain the Mean Opinion Scores (MOSs). In order to produce meaningful MOS values, 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 [6] and [7].

2.1. Test video sequences

In our subjective evaluation campaign we considered six video sequences at CIF spatial resolution (352×288 pixels), namely *Foreman*, *Hall*, *Mobile*, *Mother*, *News* and *Paris*. All the original sequences are available in raw progressive format at frame rate of 30fps. These sequences have been selected since they are representative of different levels of spatial and temporal complexity. The analysis of the content has been performed by evaluating the Spatial Information (SI) and Temporal Information (TI) indexes on the luminance component of each sequence as indicated in [8]. Additionally,

*The presented work was developed within VISNET II, a European Network of Excellence (http://www.visnet-noe.org), funded under the European commission IST FP6 programme.

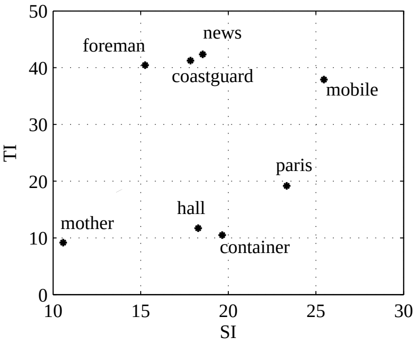


Fig. 1. Spatial Information (SI) and Temporal Information (TI) indexes of the selected video sequences [8].

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 intra mode decision	Enabled

two other sequences, namely *Coastguard* and *Container*, have been used for training the subjects, as detailed in subsection 2.3. The values of the SI and TI indexes for all the sequences are indicated in Figure 1.

Table 1 illustrates the parameters used to generate the compressed bitstreams by H.264/AVC encoding. We adopted the H.264/AVC reference software, version JM14.2, which is available for download at [9]. We encoded all sequences using the H.264/AVC High Profile to enable B-pictures and Context Adaptive Binary Arithmetic Coding (CABAC) for coding efficiency. For each sequence, 298 out of 300 frames were encoded. In fact, due to the selected GOP structure, the last two B pictures are not encoded by the reference software. Each frame is divided into a fixed number of slices, where each slice consists of a full row of macroblocks. Rate control has been disabled since it introduced visible quality fluctuations along time for some of the video sequences. Instead, a fixed Quantization Parameter (QP) has been carefully selected for each sequence so as to ensure high visual quality in the absence of packet losses. The achieved rate-distortion performance for each of the tested sequences are reported in Table 2. Briefly, we tuned the QP for each sequence in order not to exceed a bitrate of 600 kbps which can be considered an upper bound for the transmission of CIF video contents over IP networks. Each tested sequence has been visually inspected in order to see whether the chosen QPs minimized the blocking artifacts induced by lossy coding.

Table 3. Details of LCD display devices used to perform the test activity.

	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

For each of the six original H.264/AVC bitstreams corresponding to the test sequences, we generated a number of corrupted bitstreams, by dropping packets according to a given error pattern. The software that corrupts the coded contents is depicted in [10]. The coded slices belonging to the first frame are always not corrupted since they contain header information as the Picture Parameter Set (PPS) and Sequence Parameter Set (SPS). Conversely, the remaining slices are corrupted by discarding them from the coded bitstream. 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 [11]. We tuned the model parameters to obtain an average burst length of 3 packets, since it is characteristic of IP networks [12]. The two state Gilbert's model generates, for each PLR, an error pattern. Different channel realizations for each PLR are obtained by starting to read the relative error pattern at a random point. We selected two channel realizations for each PLR, for a total of 12 realizations per video content, in order to uniformly span a wide range of distortions, i.e perceived video quality, while having a dataset of reasonable dimension.

In particular, the realizations to be included in the test material have been carefully selected by applying the following steps : 1) produce for each PLR and content a set of 30 realizations and compute the corresponding PSNR values (i.e. mean PSNR values computed over the frames for each video sequence) 2) plot, for each content separately, the histograms of the PSNR values corresponding to a total of 180 realizations (i.e. 30 realizations $\times 6$ PLRs) 3) for each PLR, select on the histogram one of the most probable PSNR values so that the entire range of PSNR values is uniformly spanned 4) for each selected PSNR value, choose two corresponding realizations 5) visually check all the selected realizations to verify whether the same 5 levels of perceived quality, described in subsection 2.3, are uniformly spanned across all the different contents.

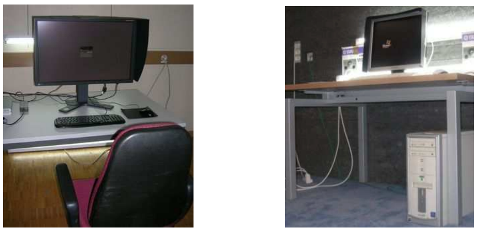
Each bitstream is decoded with the H.264/AVC reference software decoder with motion-compensated error concealment turned on [13].

2.2. Environment setup

Each test session involves only one subject per display assessing the test material. Subjects are seated directly in line with the center of the video display at a specified viewing distance, which is equal to 6-$8H$ for CIF resolution sequences, where H is the height of the video window. Accurate control and description of the test environment is necessary to assure the reproducibility of the test activity and to compare results across different laboratories and test sessions. Table 3 summarizes the crucial features of the used display devices. Pictures of the two laboratory environments are shown in Figure 2. The ambient lighting system in both the laboratories consists of neon lamps with color temperature of 6500 K.

Table 2. Test sequences

Sequence name	Spatial res.	MB/slice	Bitrate [kbps]	PSNR [db]	QPI	QPP	QPB
Foreman	CIF	22	353	34.35	32	32	32
News	CIF	22	283	37.27	31	31	31
Mobile & Calendar	CIF	22	532	28.29	36	36	36
Mother & Daughter	CIF	22	150	37.03	32	32	32
Hall monitor	CIF	22	216	36.16	32	32	32
Paris	CIF	22	480	33.64	32	32	32



(a) (b)

Fig. 2. EPFL (a) and PoliMI (b) test spaces.

2.3. Subjective evaluation procedure

In our subjective evaluation we adopt 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 includes a reference version of each video sequence, which in this case is the packet loss free sequence, as a freestanding stimulus for rating like any other.

Each sequence is 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 $[0-5]$, shown in Figure 3. Note that the numerical values attached to the scale were used only for data analysis and were not shown to the subjects.

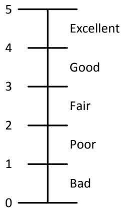


Fig. 3. Five point continuous quality scale [8].

Each subjective experiment includes the same number of 83 video sequences: 6×12 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 experiment to stabilize observers' opinion. The dummy presentations consist in 5 realizations, corresponding to 5 different quality levels, selected from the *Mobile*, *Foreman*, *Mother*, *News* and *Hall* video 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, i.e. *Coastgaurd* and *Container*, and 5 realizations of each, representatives of the score labels depicted in Figure 3. During the display of each training sequence, the trainer explains the meaning of each label, as summarized in the written 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:"

- *Excellent*: "the content in the video sequence may appear a bit blurred but no other artifacts are noticeable (i.e. it is present only the lossy coding noise)".
- *Good*: "at least one noticeable artifact is detected in the entire sequence".
- *Fair*: "several noticeable artifacts are detected, spread all over the sequence".
- *Poor*: "many noticeable artifacts and strong artifacts (i.e. artifacts which destroy the scene structure or create new patterns) are detected".
- *Bad*: "strong artifacts (i.e. artifacts which destroy the scene structure or create new patterns) are detected in the major part of the sequence".

Thus, the schedule of the experiment is the following:

- Subject training phase (approx. 5 min)
- Break to allow time to answer questions from observers
- Test phase (approx. 20 min):
  - Assessment of 5 dummy sequences
  - Assessment of 78 sequences

Twenty-three subjects and seventeen subjects participated in the tests at PoliMi and EPFL, respectively. 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 PhD students working in fields related to image and video processing, some were naive subjects.

3. SUBJECTIVE DATA PROCESSING

The raw subjective scores have been processed in order to obtain the final Mean Opinion Scores (MOS) shown in Figures 5-10, according to the steps described in Figure 4. The results of the two laboratories have been processed separately but applying the same procedure. First an ANalysis Of VAriance (ANOVA) has been performed in order to understand whether a normalization of the scores would be needed. The results of the ANOVA have shown that the difference in the subjective rates means from subject to subject was large, i.e. there were significant differences between the ways subjects used the rating scale. Thus, a subject-to-subject correction was applied, by normalizing all the scores according to offset mean correction [14]. Finally the screening of possible outlier subjects has been performed considering the normalized scores, according to the guidelines described in Section 2.3.1 of Annex 2 of [7]. Four and two outliers were detected out of 23 and 17 subjects, from the results produced at PoliMI and at EPFL, respectively. Discarding the outliers, the MOS has been computed for each test condition, together with the 95% confidence interval. Due to the limited number of subjects, the 95% confidence intervals (δ) for the mean subjective scores have been computed using the Student's t-distribution, as follows:

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

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, being N 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. It is assumed that the overlap of 95% confidence intervals provides indication of the absence of statistical differences between MOS values.

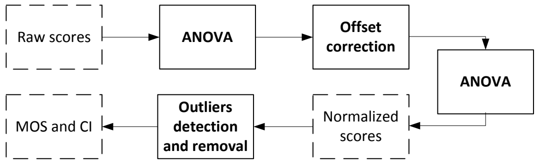


Fig. 4. Flow chart of the processing steps applied to the subjective data in order to obtain the MOS values.

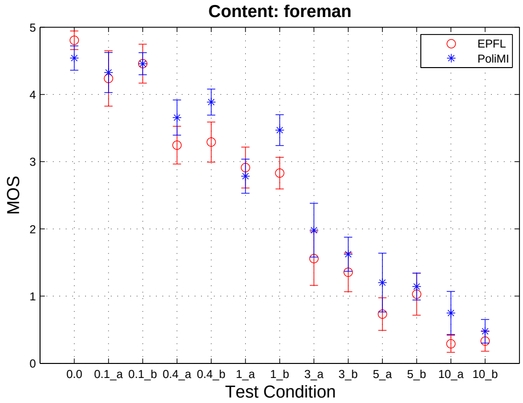


Fig. 5. MOS values and 95% confidence interval obtained by the two laboratories, for the content Foreman.

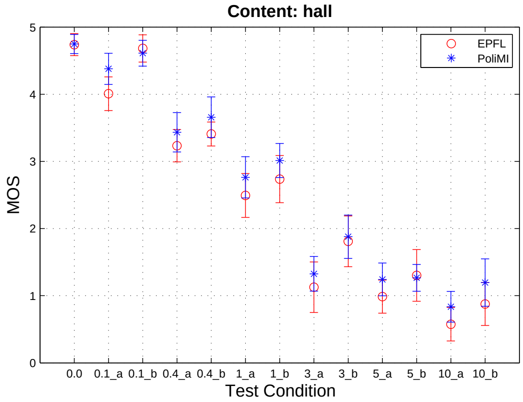


Fig. 6. MOS values and 95% confidence interval obtained by the two laboratories, for the content Hall.

4. RESULTS AND DISCUSSION

Figures 5-10 show, for each video content, the MOS values obtained after the processing applied to the subjective scores. In these figures both the MOS values collected at PoliMI and at EPFL are reported, together with their confidence intervals. Additionally, Figures 11-13 show the scatter plots between the MOS values collected at PoliMI and EPFL, together with the resulting Pearson and Spearman correlation coefficients.

As a general comment, the MOS plots clearly show that the experiment has 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.

As it can be noticed from the plots, there exists a good correlation between the data collected by the two institutions. Thus, the results can be considered equivalent and used together or interchangeably. Nevertheless, the scatter plots show as the data from PoliMI are usually slightly shifted towards better quality levels, when compared to the results obtained at EPFL. This more optimistic trend of one set of results over the other could be explained by different dot pitch values of the displays used in the two laboratories, which could mask the impairments differently. Alternatively, the shift could be due to the separate processing applied to the raw data of the two laboratories. Currently an investigation is in progess in order to better understand these aspects of the obtained results.

Finally, regarding the trend of the MOS values, it is interesting to notice as the artifacts introduced by same PLR values can be differently masked, according to the spatial and temporal complexity of the content. For example, considering the *Mother* content, the subjects clearly distinguished the quality level of the packet loss rate free sequence from the quality level of the 0.1% PLR realizations. This can be explained by the fact that this content has the lowest values of SI and TI indexes. Thus, compared to other contents, the masking effect for low PLRs is reduced.

5. CONCLUSION

In this paper the procedure followed in order to produce a publicly available dataset of subjective results for 78 CIF video sequences has been described in details. The results of the subjective tests performed in two different laboratories show high consistency and correlation. The test material (including the original uncompressed test and training material and the H.264 coded streams before and after the simulation of packet losses), the error-prone network simulator

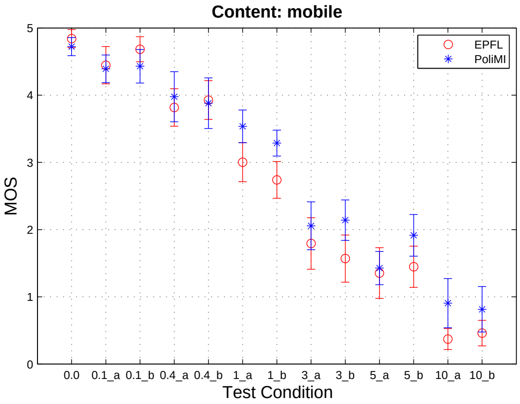


Fig. 7. MOS values and 95% confidence interval obtained by the two laboratories, for the content Mobile.

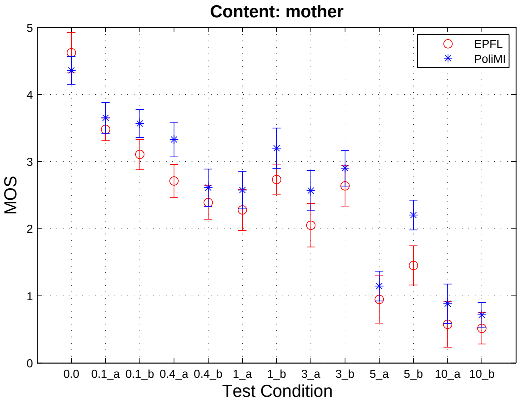


Fig. 8. MOS values and 95% confidence interval obtained by the two laboratories, for the content Mother.

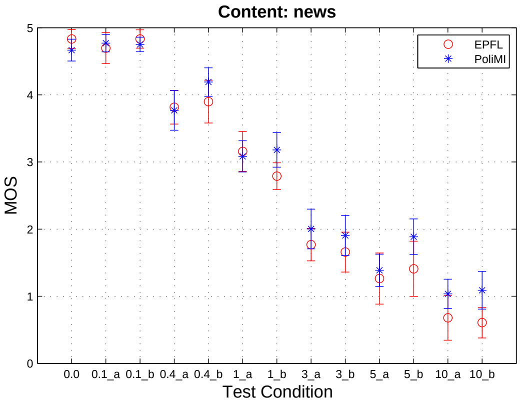


Fig. 9. MOS values and 95% confidence interval obtained by the two laboratories, for the content News.

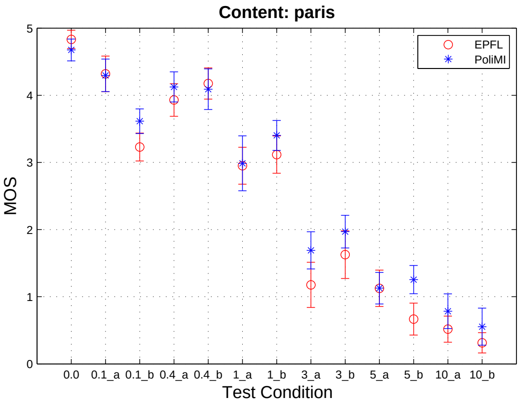
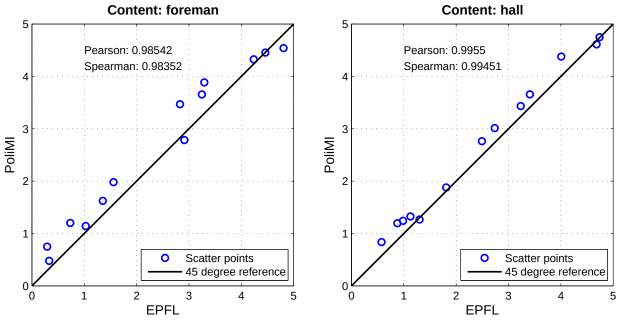
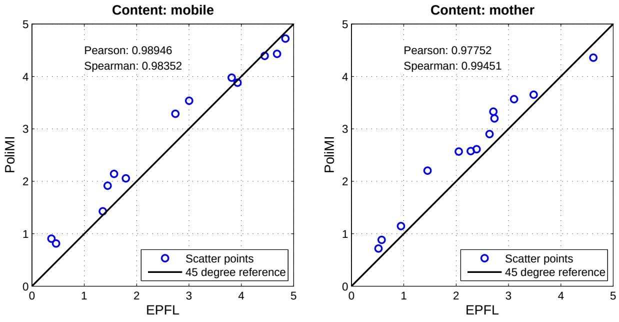


Fig. 10. MOS values and 95% confidence interval obtained by the two laboratories, for the content Paris.



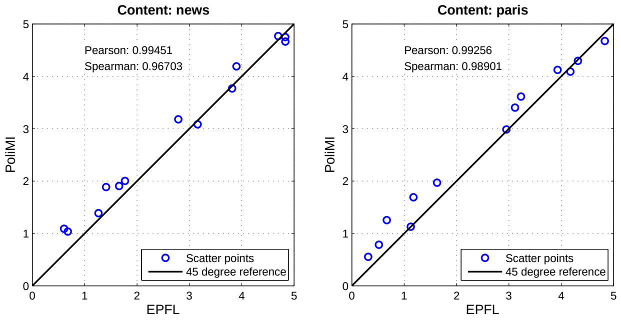
(a) (b)

Fig. 11. Scatter plot between the MOS values collected at PoliMI and EPFL for the content *Foreman* (a) and *Hall* (b), together with the resulting Pearson and Spearman correlation coefficient.



(a) (b)

Fig. 12. Scatter plot between the MOS values collected at PoliMI and EPFL for the content *Mobile* (a) and *Mother* (b), together with the resulting Pearson and Spearman correlation coefficient.



(a) (b)

Fig. 13. Scatter plot between the MOS values collected at PoliMI and EPFL for the content *News* (a) and *Paris* (b), together with the resulting Pearson and Spearman correlation coefficient.

and the H.264 decoder used in our study, the raw subjective data, the files used to process them, and the final MOS data, are available at *http://mmspl.epfl.ch/vqa*. Future works will include extension of the study to 4CIF and HD resolution data, as well as increase in the number of subjects. Finally, other test methodologies, like the continuous quality evaluation, will be taken into account.

6. REFERENCES

[1] K. Stuhlmüller, N. Färber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 1012–1032, June 2000.

[2] N. Färber, K. Stuhlmüller, and B. Girod, "Analysis of error propagation in hybrid video coding with application to error resilience," in *IEEE International Conference Image Processing*, Kobe, Japan, October 1999.

[3] I. E. G. Richardson, *Video Codec Design*, John Wiley & Sons, 2002.

[4] M. Claypool and J. Tanner, "The effects of jitter on the perceptual quality of video," in *ACM Multimedia*, Orlando, FL, USA, November 1999.

[5] Yuan-Chi Chang, Thom Carney, Stanley A. Klein, David G. Messerschmitt, and Avideh Zakhor, "Effects of temporal jitter on video quality: assessment using psychophysical and computational modeling methods," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, B. E. Rogowitz and T. N. Pappas, Eds., July 1998, vol. 3299 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp. 173–179.

[6] "Vqeg hybrid testplan, version 1.2," ftp://vqeg.its.bldrdoc.gov.

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

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

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

[10] M. Luttrell, S. Wenger, and M. Gallant, "New versions of packet loss environment and pseudomux tools," Tech. Rep., Joint Video Team (JVT), October 1999.

[11] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell System Technical Journal*, vol. 39, pp. 1253–1266, September 1960.

[12] 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.

[13] 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), September 2003.

[14] E. Drelie Gelasca, *Full-reference objective quality metrics for video watermarking, video segmentation and 3D model watermarking*, Ph.D. thesis, EPFL, September 2005.