QUALITY ASSESSMENT FOR H.264 CODED LOW-RATE AND LOW-RESOLUTION VIDEO SEQUENCES

Olivia Nemethova, Michal Ries and Markus Rupp
Institute for Communications and RF Engineering
Vienna University of Technology
Gusshausstr. 25/389
Vienna, Austria
email: {onemeth, mries, mrupp}@nt.tuwien.ac.at

Eduard Siffel
Institute for Telecommunications
Slovak University of Technology in Bratislava
Ilkovicova 3
Bratislava, Slovakia
email:siffel@ktl.elf.stuba.sk

ABSTRACT
This article concentrates on a quality assessment for H.264 coded low-rate and low-resolution video sequences which are in particular of interest for mobile communication. The choice of appropriate setup for tests is discussed. The focus is given on the influence of the sequence character (spatial and temporal information). The data set is compared to various known video quality metrics. It is shown on the obtained data set, that it is not possible to separate the dynamic and static parameters without considering the character of the sequence and thus to create a universal metric.

KEY WORDS
H.264, subjective perceptual video quality, spacial information, temporal information.

1 Introduction

The deployment of packet-oriented wireless networks offers new mobile multimedia applications like MMS, video streaming and video-conferencing. Such applications introduce several new challenges as they are delay-sensitive and use relatively high bandwidth. Typical end-terminals for such services are the mobile phones, using low QCIF (144×176 pixel) image resolution. Due to the bandwidth limitations of wireless transmissions, it is necessary to compress the video stream before the transmission by means of using the lossy compression algorithms or/and frame rate reduction, introducing particular quality degradation, that can be observed as a distortion to the temporal continuity or static picture quality. Several tradeoffs are required between the quality and the amount of the resources needed for the various video application.

While many researchers[1, 2] focus on relative simple but objective measures like the Peak-to-Signal to Noise Ratio (PSNR), newer results decide which degradation is (still) acceptable for the user by assessing and estimating his subjective perceptual quality evaluation, given by a so-called mean opinion score (MOS). MOS is metric well-known from the subjective perceptual quality evaluation of audio sequences. To obtain such MOS for the one-directional transmission of video sequences, several human observer test methods are described in [3].

Performing a subjective video quality survey requires much effort making it impossible to perform it anytime and anywhere. Therefore, there are several metric proposals (e.g. [4, 5]) how to extract MOS values from the video sequence parameters set at the sender or calculated using the model of human visual perception after the reception at the receiver.

However, subjective quality evaluation is a psycho-visual experiment and thus the results strongly depend on the type and character of the sequence itself. The intention of this paper is to demonstrate the dependency of the MOS on the sequence character by means of a survey, and to compare obtained results with known metrics.

In Section 2 the sequences selected for evaluation are described as well as the setup of the survey which we performed to obtain MOS values. In Section 3 some known metrics for video quality are applied to our set of data and evaluated. The results are further interpreted in Section 4. Focus is given on the video sequence characteristics. Section 5 contains the conclusions and some final remarks.

2 VIDEO QUALITY SURVEY

For the tests we selected four video sequences each of ten-second duration with QCIF resolution. Two of them (akiyo, foreman) are well-known professional test sequences obtained by a static camera. In the akiyo sequence a female moderator is reading news only by moving her lips and eyes. The foreman sequence contains a monologue of a man moving his head dynamically and at the end of the sequence there is a contiguous scene change. Soccer and panorama are both sequences with permanent camera movement. Soccer is a professional sequence; the entire picture is moving - the players and ball in a fast way, the background rather slowly. Panorama is a non-professional sequence, containing uniform but smooth and relatively slow movements of the scene. Snapshots of these sequences are depicted in Figure 1.

We used all possible nine combinations of bit rates 128kbps, 64kbps, 32kbps and frame rates 15fps, 10fps,

Figure 1. Video test sequences used in the survey: akiyo, foreman, soccer, panorama.

5fps shown in the following table as well as a non-compressed sequence.

bit rate [kbps]	frame rate [fps]
128	15
128	10
128	5
64	15
64	10
64	5
32	15
32	10
32	5

We have chosen the H.264/AVC codec (more details about it can be found in [6]), a recent video coding standard of the ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group. This codec contains new technical features that improve its compression performance while keeping the same quality.

To evaluate the subjective perceptual quality, we worked with 30 unpaid voluntary test persons. The group of subjects was chosen as diverse as possible, ranging different ages, skills and backgrounds. After the test was performed, we asked the subjects to further fulfil a small questionnaire in order to obtain an information about their age, sex, education and experience with imaging.

The tests were performed according to [3], using a CRT screen; the QCIF picture located in the middle of the gray background. We performed absolute category rating tests. At the beginning a trial run with similar video sequences (ice-hockey, salesman, winter-nature) and typical coding artifacts was performed. Test subject were asked to evaluate the overall quality, static quality and the temporal continuity on the scale with nine grades (1-bad, 3-poor, 5-fair, 7-good, 9-excellent). Different sequences were presented in an arbitrary order, with additional condition that the same sequence (even differently degraded) did not appear in succession. If we had presented always the same sequence with different degradations first and then another sequence etc., we would have obtained only the subject response on different encoding combinations. To obtain information on how the subjects evaluate different sequences relatively to each other, we had to alternate the sequences.

A component analysis has been performed in order to compare our results with those in [4] and to find correlations of our dynamic (temporal continuity) and static (static quality) subjective parameters with the subjective quality parameter overall quality. The result of such analysis is shown in Figure 2.

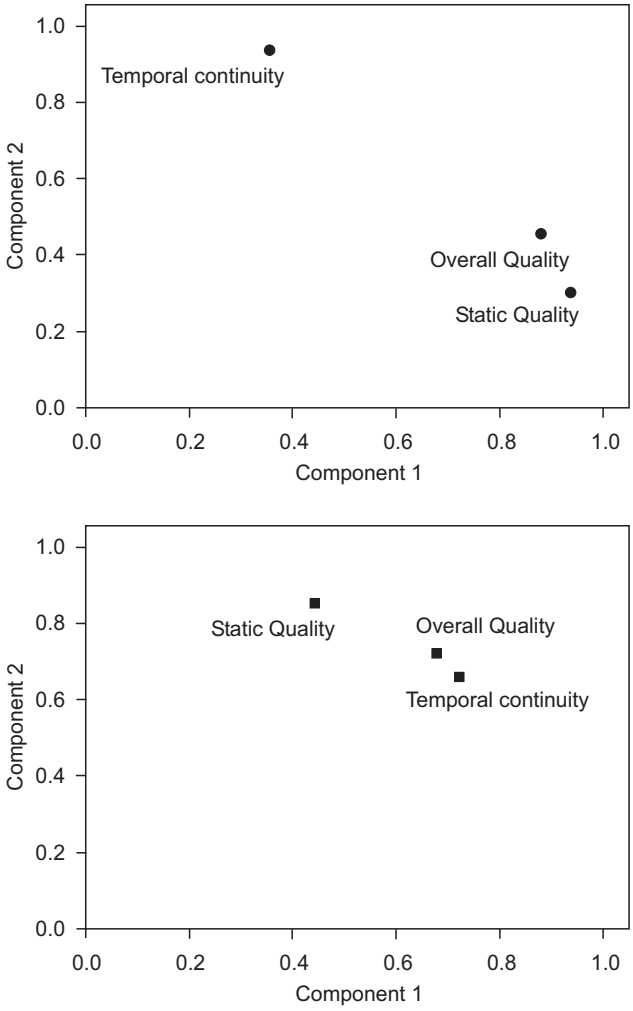

Figure 2. Component analysis for subjective perceptual parameters overall quality, static quality and temporal quality: on top for all sequences without soccer, on bottom for soccer

It can be seen that for the soccer sequence the overall quality correlates more with the subjective parameter tempo-

ral continuity, which should reflect the frame rate. For the other sequences the overall quality is more correlated with the subjective static quality parameter (at most in case of panorama sequence), which should reflect the PSNR or any other objective static parameter. These results let one suspect that the character of a sequence may play crucial role in the subjective evaluation.

3 QUALITY METRICS EVALUATION

For a picture with luminance quantized by q bits, the PSNR in [dB] is defined as follows:

$$\text{PSNR} = 10\log_{10}\frac{(2^q)^2}{\frac{1}{MN}\sum_{x=1}^{N}\sum_{y=1}^{M}\left[a(x,y)-b(x,y)\right]^2}, \quad (1)$$

where a is an $N \times N$ observed picture and b is $M \times M$ original picture, x and y being the pixel coordinates.
This metric may be suitable for still pictures but video has additionally a temporal dimension that influences the subjective quality as well. Therefore, in [4] the following MOS prediction metric Q_m has been proposed:

$$Q_m = -0.45\text{PSNR} + 17.9 - 0.1(\text{FR} - 5), \quad (2)$$

where FR is a frame rate of the video sequence and the constant coefficients were interpolated by evaluating of the data set obtained in a survey. Please note, that this prediction considers a five grade MOS scale, the best grade being 1 (opposit to the usual MOS scales, where the higher number corresponds to higher quality [3]). We therefore adapted this metric to our nine grade scale.
For the Q_m metric, the frame rate only results in an offset of the linear mapping between the PSNR and the MOS.
In Figure 3 the relation of the above mentioned metrics with the overall quality parameter evaluated by human observers is shown for the sequences akiyo, foreman and soccer. One can see, that our results in the given interval do not fit well neither the PSNR metric, nor the Q_m metric. We obtained a much better MOS fit by

$$Q_{\text{fit}} = a - \frac{b}{\text{PSNR}}, \quad (3)$$

with the coefficients $a = 14.2$ and $b = 280.5$. The correlation coefficient for the data in given interval was $r = 0.909$. The fit is only to show, that if there is a relation of MOS and PSNR, it will not be linear in the relevant range. To obtain a really consistent metric more data would be required.
There are several reasons for the result, that the linear metric proposed in [4] does not fit our data at all. One of them is the choice of a different codec. Usage of different codecs or even different implementations of the same codec, results in different PSNR for the same bit rate and frame rate. Anyhow, it is not suitable to use the linear metric, as MOS is a finite scale. Another reason is the character of the sequence, which seems to have an essential influence on the subjective evaluation as will be shown in the next section.

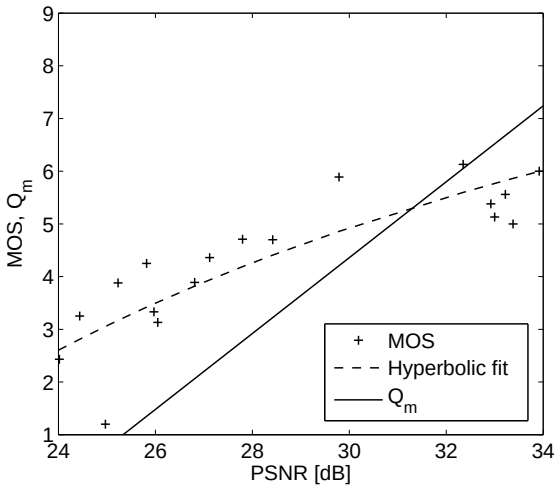

Figure 3. Mean opinion score evaluation of the overall quality and Q_m prediction over PSNR.

Averaging over the sequences does not bring relevant results, as the subjective perception differs for the same objective static and dynamic parameters if the sequences have different character.

4 SEQUENCE CHARACTERISTICS

In Figure 4 the relation between PSNR and MOS is presented for our four sequences separately.

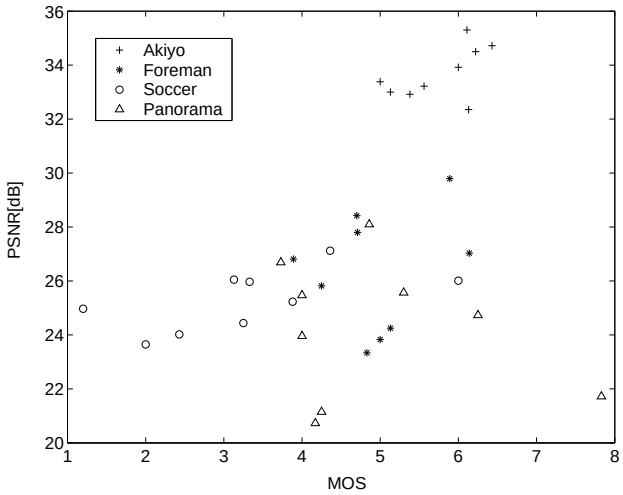

Figure 4. Relation between MOS and PSNR for different video sequences.

The depicted relation in Figure 4 looks completely different for different sequences. It can be concluded again that the PSNR is not a suitable metric for video quality as there

is no clear mapping between PSNR and MOS without considering at last a particular frame rate.

In [3] the measure of spatial and temporal perceptual information are used to characterize a video sequence. The spatial perceptual information measurement (SI) is based on the Sobel filter, that is applied to each luminance frame F_n at time instance n. After that the standard deviation over the pixels is computed. The maximum value within the whole sequence represents the spatial information:

$$SI = \max_{time} \{\text{std}_{space}[\text{Sobel}(F_n)]\}. \quad (4)$$

The temporal perceptual information measurement is based upon the motion difference feature. For every time instance n, the luminance pixel values difference is counted: $M_n(i,j) = F_n(i,j) - F_{n-1}(i,j)$. TI is computed as a maximum over time of the standard deviation over space:

$$TI = \max_{time} \{\text{std}_{space}[M_n(i,j)]\}. \quad (5)$$

In the following table SI and TI for our four sequences are listed.

sequence	SI	TI
akiyo	79.4	5.2
foreman	105.3	36.5
soccer	85.8	22.9
panorama	138.1	22.1

Sequence akiyo can be compressed easily as it contains low amounts of both spacial and temporal information. Therefore, for the same bit rates, we obtain higher PSNR than for another sequences. In Figure 5 it can be seen, that also the MOS for the akiyo sequence does not vary much. Interesting is the fact, that there is almost no difference in subjective quality between the combinations $32/5$ and $128/15$. For the sequence foreman the MOS is similar to the akiyo sequence. The users evaluate this sequence more critically than the sequence akiyo. They are more sensitive to the frame rate. The entire sequence is of more dynamic nature than akiyo, but not as contiguous dynamic as for example soccer, although the metric is higher for foreman. This is caused by the late part of the foreman sequence with the fast (but still contiguous) scene change.
In Figure 6 the results for the soccer and panorama sequences are presented. The test subjects were especially critical to the soccer sequence. This is caused by the fact, that in this sequence the most important information is concentrated in the smallest and most dynamic object in the picture - in the ball. Also important are the lines on the playground. As this sequence is very dynamic, the user are sensible to the frame rates. Insofar as the ball and the player can be seen, the user prefer worse static quality rather than low frame rate. Very interesting is the fact, that the combination 32/15 seemed to be liked better by the user than the combination 128/15. As H.264 uses the in-the-loop deblocking filter, the playground in the combination 32/15 is

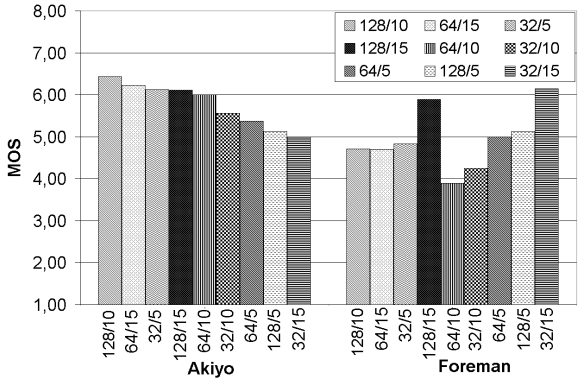

Figure 5. MOS for all tested codec combinations frame rate/bit rate for sequences akiyo and foreman.

rather blurred, but the players and ball are still well visible. Using the frame rate 15, the movement seems to be more contiguous for the blurred area than for the area with more spatial details. Even more surprising is the MOS for panorama sequence. In this case, the most important parameter seems to be the static quality - the quality of particular 'still' pictures of the sequence. The very best result we obtained for the coding combination 128/5, which let suspect that because of uniformity of the movement, the human eye can better approximate in the temporal domain than in the spacial.

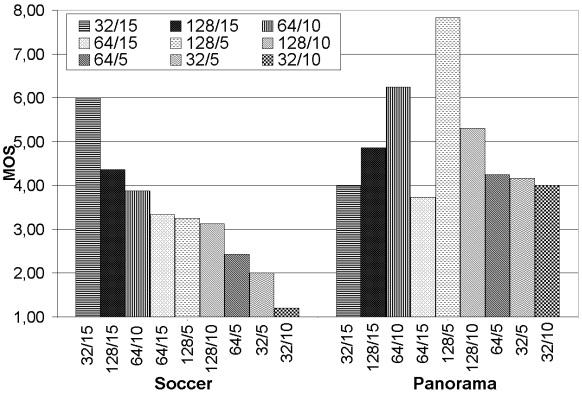

Figure 6. MOS for all tested codec combinations frame rate/bit rate for sequences soccer and panorama.

In Figure 6 the MOS values are averaged over all users. As already mentioned, we performed the tests with very diverse subjects. Therefore after averaging, the highest MOS value does not reach grade 8 of the 9 grade scale and most of the best evaluations are around grade 6. There was an apparent difference between the evaluations of the people usually working with computers or having some experience with imaging and the remaining part of the group.

In Figure 7 and 8 the MOS for overall quality in dependence on the frame rate and the bit rate can be seen for the very different sequences akiyo and soccer, respectively. From these figures it can be seen, that not only different frame rates, but also the sequence character influences the shape of the curve (and not only the offset as in Q_m).

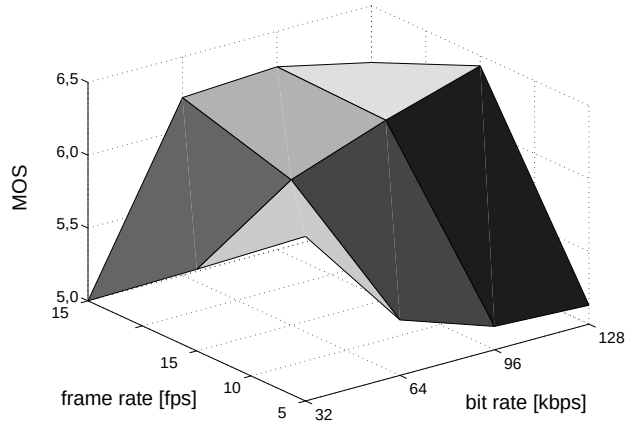

Figure 7. Relation between the frame rate, bit rate and MOS for akiyo sequence.

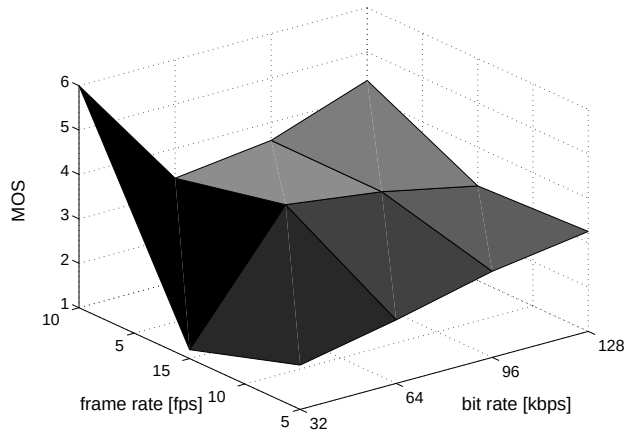

Figure 8. Relation between the frame rate, bit rate and MOS for soccer sequence.

In order to propose a universal prediction formula for the subjective quality, we need to take into account a parameter that corresponds to the rate reduction, a parameter corresponding to the PSNR reduction **as well as** parameter(s) corresponding to the sequence character. For such a step much more test data are required.

5 CONCLUSION

We presented a comparison of the subjective perceptual quality evaluation results, obtained in a survey with some objective video parameters and some known metrics. Test sequences and subjects were selected to be as representative as possible. Our results confirmed that PSNR alone as a static parameter does not reflect the subjective perceptual quality evaluation, and moreover that even the combination of a static parameter with a dynamic parameter, given for example by frame rate is not a suitable metric to predict the subjective evaluation. The reason for such conclusion is the very different subjective quality evaluation under the same conditions for the video sequences with different character. For some sequences we obtained even better results for low frame rate (5fps), a consequence of the particular human eye interpolation. To be able to propose a relevant metric, much more test data are required, based on many different sequences and for more possible coding combinations. For particular sequence types the metric is expected to look very like the presented surface graphs. In this paper we analyzed the very simple possibility to base the metric on the coding parameters. We showed, that this straight forward method only suites for similar sequence types.

6 ACKNOWLEDGEMENT

We would like to thank all voluntary test subjects.

References

[1] P. Bucciol, E. Masala, J.C. De Martin, ”Perceptual ARQ for H.264 Video Streaming over 3G Wireless Networks,” Proc. of ICC 2004, Paris, France, June 2004.

[2] G. Cheung, C.N. Cuah, D.J. Li, ”Optimizing Video Streaming Against Transient Failures and Routing Instability,” Proc. of ICC 2004, Paris, France, June 2004.

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

[4] G. Hauske, T. Stockhammer, R. Hofmaier, ”Subjective Image Quality of Low-rate and Low-Resolution Video Sequences” , Proc. International Workshop on Mobile Multimedia Communication, Munich, Germany, Oct. 5-8, 2003.

[5] S. Winkler, F. Dufaux, ”Video Quality Evaluation for Mobile Applications”, Proc of SPIE Conference on Visual Communications and Image Processing, Lugano, Switzerland, vol. 5150, pp. 593-603, July 2003.

[6] T. Wiegand, G.J. Sullivan, G. Bjontegaard, G.; A. Luthra, ”Overview of the H.264/AVC Video Coding Standard”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 560-576, July 2003.