Subjective Evaluation of Video Quality for H.264 Encoded Sequences

Olivia Nemethova*, Michal Ries*, Eduard Siffel**, Markus Rupp*

*Institute of Communications and Radio Frequency Engineering
Vienna University of Technology
Gusshausstr.25, A-1040 Vienna, Austria
{*onemeth, mries, mrupp*}*@nt.tuwien.ac.at*
** Faculty of Electrical Engineering and Information Technology
Slovak Technical University in Bratislava
Ilkovicova 3, 812 19 Bratislava, Slovakia
siffel@ktl.elf.stuba.sk

ABSTRACT

The newest video coding standard H.264/AVC has achieved significant improvements by means of several new technical features, allowing for better adaptation to the underlying network. In this article we present an evaluation of user subjective quality for low bit rate video suitable for transmission over wireless systems. We compare the existing objective quality metrics, show their limitations regarding the mean opinion score (MOS) prediction, and discuss how to obtain a relevant metric according to the survey data.

1 INTRODUCTION

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). 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 counted after the reception.

The H.264/AVC codec (more details about it can be found in [7]) is 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, which make it in particular suitable for wireless networks. Blocking is generally considered to be one of the most visible artifacts when using block-based coding. For this reason, H.264/AVC defines an adaptive in-loop deblocking filter, where the strength of filtering is controlled by the values of several syntax elements. The blockiness is reduced without much affecting the sharpness of the content. Consequently, the subjective quality is significantly improved. Furthermore, H.264 uses small block-size transformations allowing the encoder to represent signals in a more locally-adaptive manner, which reduces artifacts known as 'ringing'.

The intention of this paper is to provide a comparison of new subjective perceptual quality assessment results for H.264 encoded video with objective metrics known so far. 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 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, 5fps shown in the following table as well as a non-compressed sequence.

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. 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. Test subjects 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,

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

with additional condition that the same sequence (even differently degraded) did not appear in succession. 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. A component analysis has been performed in

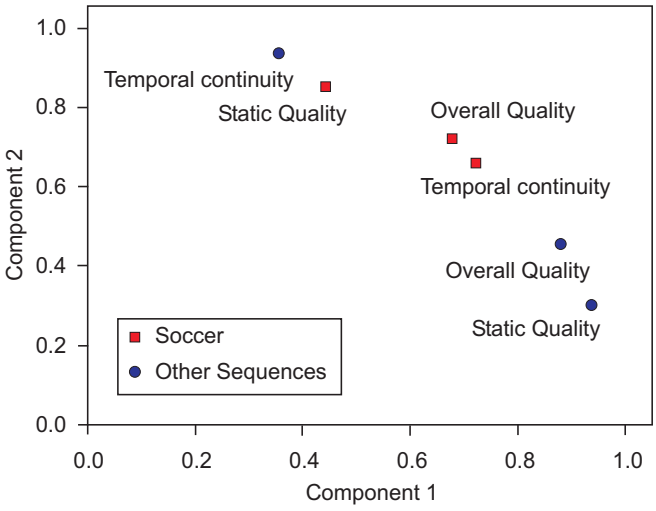

Figure 2: Component analysis for subjective perceptual parameters

order to compare our results with those in [4] and to find correlations of our dynamic and static subjective parameters with the overall quality. The result of the analysis is shown in Figure 2.
For the soccer sequence the overall quality correlates more with the subjective parameter temporal continuity, which should reflect the frame rate. For the other sequences the overall quality is more correlated with the subjective static quality parameter, which should reflect the PSNR or any other objective static parameter.

3 QUALITY METRICS EVALUATION

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

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

where a is an $N \times N$ observed picture and b is the $M \times M$ 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:

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

where FR is the frame rate of the video sequence. Please note, that this prediction considers a five grade MOS scale, the best grade being 1 (opposite 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

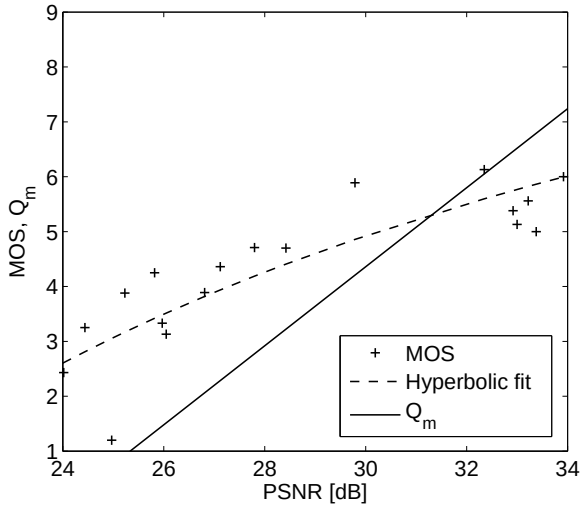

Figure 3: Mean opinion score evaluation of the overall quality and Q_m prediction over PSNR for FR=10fps.

do not fit well neither the PSNR metric, nor the Q_m metric. We obtained a much better MOS fit by

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

with the coefficients $a = 14.2$, $b = 280.5$ and the correlation coefficient $r = 0.909$. 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.

The linear metric proposed in [4] does not fit our data at all. One possible reason 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. Another reason is the averaging over the different types of sequences as the subjective visual perception evaluation differs considerably for the same objective static and dynamic parameters in that case.

Another approach for the MOS prediction has been adopted in [5] and [6]: a perceptual distortion metric is based on a model for human visual perception. Perceptual distortion results from the degradation of quality using high data rate reduction. Human eyes are especially sensitive to the distortion of edges, blockiness, ringing effect or jerkiness. According to the mentioned authors, the metric prediction may be obtained by considering these effects as a part of metrics, finding the weight factors by means of a survey.

Because of the deblocking filter and smaller transform blocks used in H.264, the measures as blockiness or ringing effect are not as applicable as for the previous codecs. Bluriness might apply better for this case. According to [5] we counted the bluriness as an average thickness of the edges. The mapping between the MOS and counted bluriness can be seen in the Figure 4 for different frame rates. One can see, that this measure alone without considering the type of sequence does not reflect the static quality as perceived by the users. In some cases the original picture is more blurry than the compressed one, which is a result of the H.264 optimization.

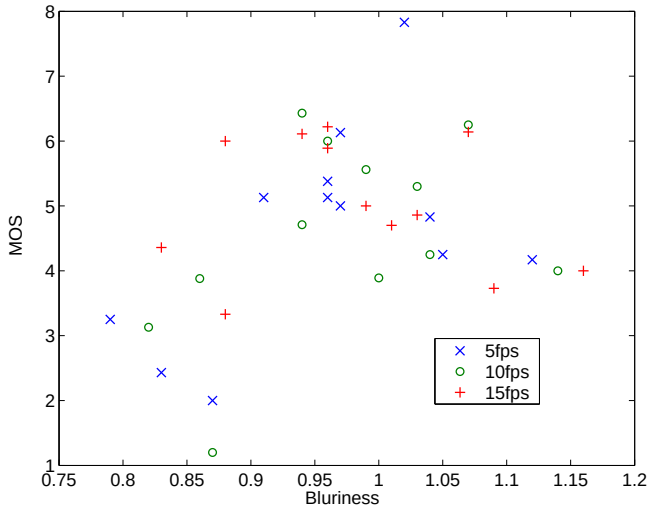

Figure 4: Relation between the bluriness metric and the MOS for tested frame rates. The bluriness is normalized to the original uncoded picture

4 INTERPRETATION of RESULTS

In Figure 5 the relation between PSNR and MOS is presented for our four sequences separately.

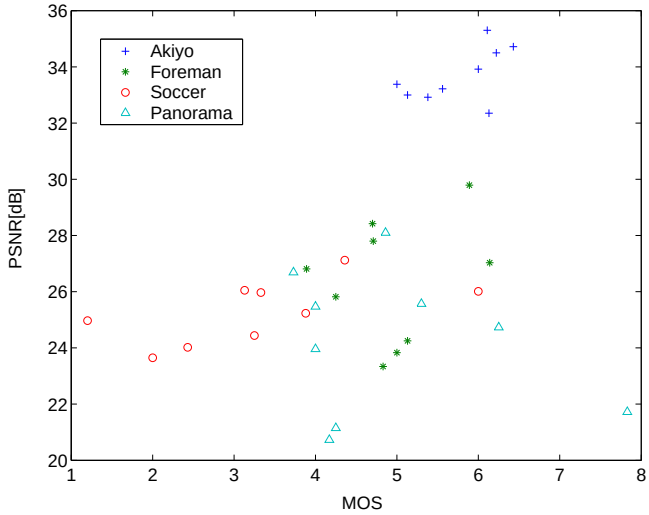

Figure 5: Relation between MOS and PSNR for different video sequences.

The depicted relation in Figure 5 looks very 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 the dynamic part of video.
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:

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

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:

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

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 small amounts of both spacial and temporal information. Therefore, for the same bit rates, we obtain higher PSNR than for another sequences. In Figure 6 it can be seen, that also the MOS for the akiyo sequence does not vary much. For the sequence foreman the MOS is similar to the akiyo sequence. The users 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.

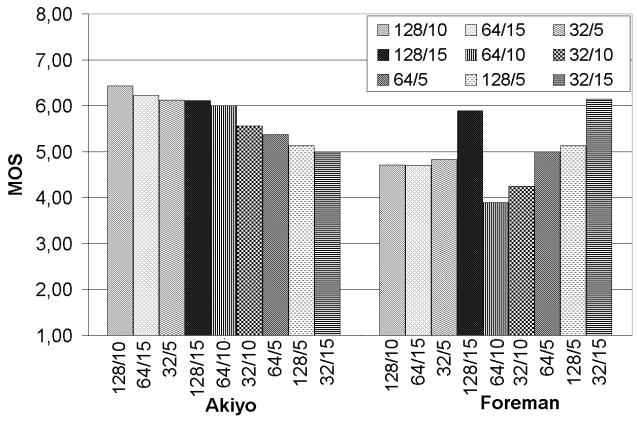

Figure 6: MOS for all tested codec combinations frame rate/bit rate for sequences akiyo and foreman.

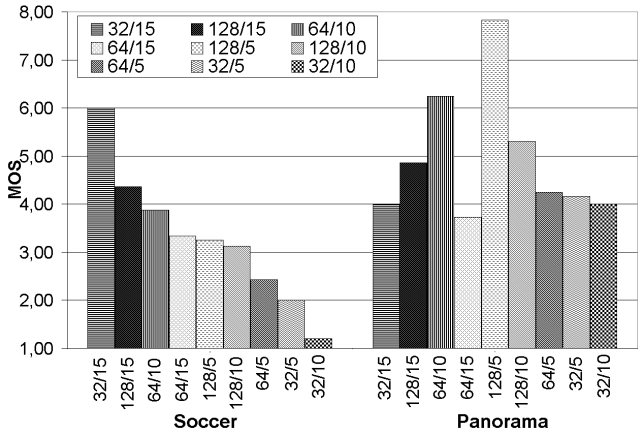

Figure 7: MOS for all tested codec combinations frame rate/bit rate for sequences soccer and panorama.

In Figure 7 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. Insofar as the ball and the player can be seen, the user prefer worse static quality rather than low frame rate. 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 rather blurred but the players and ball are still well visible. For panorama 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.

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. H.264 codec has been optimized for perceptual quality and therefore the results with the known metrics do not lead to a good fit. Our results confirmed that static parameters as PSNR or bluriness does not reflect the subjective quality evaluation, and that even the combination of such static parameter with a dynamic parameter, given for example by frame rate is not a suitable metric to predict the subjective evaluation. 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.

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. of 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] E. Ong, W. Lin, Z. Lu, S. Yao, X. Yang, F. Moschetti, "Low Bit Rate Video Quality Assessment Based on Perceptual Characteristics", Proc. of International Conference on Image Processing, 14-17 Sept. 2003, vol. 3, pp. 189-192.

[7] 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.