PSNR-Based Estimation of Subjective Time-Variant Video Quality for Mobiles

Olivia Nemethova, Michal Ries, Matej Zavodsky and Markus Rupp

Institute for Communications and Radio-Frequency Engineering,
Vienna University of Technology
Gusshausstasse 25, A-1040 Vienna, Austria
{onemeth, mries, mrupp}@nt.tuwien.ac.at

***Abstract*— With low-rate video services over packet networks, the topic of video quality measuring became even more popular. In the presence of packet loss, the video quality becomes highly time-variant. The subjective quality evaluation is in general correlated with the metrics based on the difference between the degraded sequence and the reference. However, it can differ significantly in some cases, thus it provides limited means for the appropriate quality estimation. In this work we focus on the relation between the peak to signal-to-noise ratio (PSNR) and the mean opinion score for the video streams with time variant quality. We study simple human perception rules and use them to correct the PSNR accordingly, to achieve suitable subjective quality estimation, based on objective parameters. We focus on mobile terminals and video applications. Thus, we performed the assessments on mobile terminals using our proposed methodology. The results show considerable improvement of the estimation quality with respect to the PSNR.**

I. Introduction

Real-time packet video services in error-prone environments suffer from packet loss. Client applications at the receiver cope with the losses in different ways, using various spatial and/or temporal error concealment methods, that are usually implementation specific – not standardized. Therefore, the received video will contain some impairments which influence the end-user perceived video quality. For provisioning of quality of service it is important to monitor the network performance at the user side. One task connected with network monitoring is the estimation of the end-user perceived quality. Extensive research has been performed in this field during the last decade. In most of the cases, the focus was in broadcasting and internet applications. Therefore, also the recommendations describing the tests methodology [1] and metric synthesis [2] are based on such assumptions. However, according to what we observed performing various tests, the user evaluation of the same video shown on the PC and on the mobile phone can differ. Therefore, the metrics synthesized out of such tests may not reflect the user perception. In this paper we present the results of subjective tests performed on mobile terminals. We present and evaluate a modified test methodology. The obtained mean opinion scores (MOS) are mapped onto the peak to signal-to-noise ratio (PSNR) metrics. The PSNR does not perform well for the cases with compression impairments only [3], but shows a better performance for prevailing impairments caused by losses [2]. The match of the PSNR and MOS we improve by appropriate scaling and smoothing of the PSNR curve. We further analyze the differences between the MOS and the adapted PSNR curves and propose rule-based corrections, which considerably improves the MOS prediction performance. Several alternative approaches to the quality estimation have already been proposed. The effect of scene changes on the human perception has been studied already in [4], a general study of the viewer response to time variant video quality can be found in [5]. A set of metrics for impairment detection, based on the blockiness, blurriness, jerkiness and ringing artifacts can be found in [6], [7] and others. These metrics are specially suitable for reference-free quality estimation and to capture the impairments caused by the compression. We show that based on the widely used PSNR metric, appropriately corrected, we can obtain fairly well results.

This article is organized as follows: In Section 2 the sequences selected for evaluation are described as well as the setup of survey performed to obtain the MOS values. Section 3 describes the data analysis, while Section 4 describes the quality metric design. Results are presented and further interpreted in Section 5. Section 6 contains conclusions and some final remarks.

II. Subjective tests of time-variant quality

For the subjective video quality tests, we selected six video sequences with different content features, each having approximate duration of 2min and the QCIF resolution (144×176 pixel). Some screenshots of these video sequences are depicted in Figure 1. The ”news 1” and ”news 2” video sequences contain more different scenes with both, static and moving camera (moderator reading news, scenes illustrating the news content). The scenes are usually separated by scene cuts, in three cases by fast zooming zooming-in/zooming-out scene change. The video sequence ”soccer” is a typical soccer match, containing wide-angle panning camera as well as close-up scenes, separated from each other by cuts. The ”cartoon” sequence contains two different short cartoons. Scene separation is performed mostly by transitions, in some cases by cuts. The ”city” sequence is a picture guide over the city. Different scenes are separated by slow transitions. The clip ”traffic” is a

Fig. 1. Screenshots of some video test sequences used in the survey: ”news 2”, ”soccer”, ”cartoon”, ”city”.

static camera sequence containing views on four traffic nodes in a city, separated by cuts.

All sequences were encoded using H.263 profile 0 and level 10. For subjective quality testing we used combinations of bit rates and frame rates (7.5 fps) shown in Table I together with the number of cuts in these sequences and the packet loss probability p (memoryless channel). PSNR determines the overall degradation of the video sequence; it is averaged over time. To obtain MOS, we performed three

Video sequence	Bit Rate	Scenes	p	PSNR[dB]
”news 1”	44 kbit/s	12	0.07	15.80
”news 2”	56 kbit/s	15	0.08	22.96
”soccer”	105 kbit/s	7	0.01	36.88
”cartoon”	44 kbit/s	25	0.10	23.83
”city”	80 kbit/s	39	0.05	25.14
”traffic”	44 kbit/s	4	0.05	28.47

TABLE I

PARAMETERS OF THE TEST SEQUENCES. ADDITIONALLY, ALL VIDEO SEQUENCES WERE TESTED WITHOUT PACKET LOSS (COMPRESSED).

runs of tests with 38 paid test persons and evaluated them. The chosen group ranged different ages (between 17 and 30), sex, education and experience with image processing. The testing environment can be seen in Figure 2. In [1], there are two methods described for time variant video testing: Single Stimulus Continuous Quality Evaluation (SSCQE) and Double Stimulus Continuous Quality Evaluation (DSCQE). Neither of them suited our needs: SSCQE does not take into account the original sequence at all; DSCQE switches between the original and the degraded version which does not occur in real world. In [8] was been shown that SSCQS appropriately performed also results in a set of consistent data. We decided to adopt the SSCQE as a basis and additionally we show the people also (compressed) original sequence (without packet loss), without informing them about it. We adapted the evaluation of results correspondingly as described in Section III.

According to [1] and [2], the tests should be always performed on a PC screen, under strictly specified conditions.

Fig. 2. Testing environment: test person holding the mobile telephone, evaluating with slider.

This, however, does not correspond to our desired scenario of estimating the video quality for the applications running on the mobile terminals. The test persons are evaluating more critically if the video is played-back at the PC. Therefore, our test sequences were played-back on the UMTS mobile terminal Sony-Ericsson Z1010. The display distance was chosen by the test person according to his/her convenience. All tests were performed in the same laboratory.

The quality feedback was captured by a hardware slider [10]. The interval between each two samples was 150 ms, the quality scale was between 0 (lowest quality) and 255 (highest quality). We use this scale for MOS throughout the paper. With each test person we performed three runs: the second run took place one hour after the first run, the third run followed two weeks later.

III. DATA SET PROCESSING

The set of 114 measured time variant quality curves obtained per video sequence we first tested on the consistency by the 95% confidence interval test [1]. The resulting histogram of the 95% confidence interval δ can be seen in Figure 3. We performed 95% screening as recommended in [1] – there

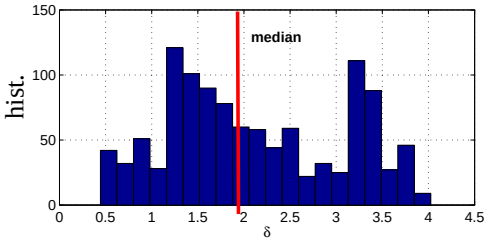

Fig. 3. Histogram of the 95% confidence interval.

were no tests we had to exclude. Thus the applied test methods provided a consistent data set.

The resulting MOS$[n]$ curve was obtained by sample-wise averaging over all curves for the given sequence. In Figure 4 the MOS$[n]$ for the sequence ”news 2” with (MOS$_{\text{deg}}[n]$) and

without ($\text{MOS}_{\text{ori}}[n]$) packet loss is shown for the frame number n. Note that the $\text{MOS}_{\text{ori}}[n]$ for the video without packet loss

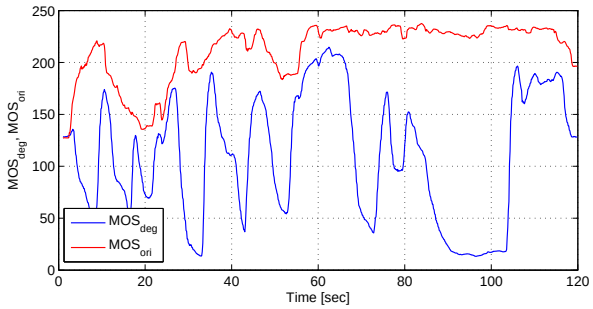

Fig. 4. MOS for the sequence "news 2" with and without packet loss.

can also vary. This is caused by the compression artifacts depending on the instantaneous sequence character and by the quality of shots. However, PSNR was shown not to reflect well the subjective quality of the compression artifacts [3]. To suppress the influence of these effects, we will further work with the corrected MOS:

$$\text{MOS}[n] = \text{MOS}_{\text{deg}}[n] + \big(\text{MAX} - \text{MOS}_{\text{ori}}[n]\big), \qquad (1)$$

where $\text{MAX} = \max_n \text{MOS}_{\text{ori}}$. The last correction we need to perform is the compensation of the time shift caused by the human and equipment reaction time. We needed to apply the shift of 150 ms the mean opinion score curve.

IV. Estimation of Subjective Quality

The mostly used reference video quality metric is the reference-based peak to signal-to-noise ratio (PSNR)

$$\text{PSNR}[n] = 10 \cdot \log_{10} \frac{255^2}{\text{MSE}[n]}, \qquad (2)$$

$$\text{MSE}[n] = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{c=1}^{C}[\mathbf{F}_n^{(c)}(i,j) - \mathbf{R}_n^{(c)}(i,j)]^2}{N \cdot M \cdot C}, \qquad (3)$$

where $\text{MSE}[n]$ denotes the mean square error of the nth frame $\mathbf{F}_n$ compared to the frame $\mathbf{R}_n$ of the reference sequence. The resolution of the frames is $N \times M$, indexes i and j address particular pixel values within the frame, C is the number of the color components and c is an index to address them. The reference sequence can be the original compressed or non-compressed sequence. We use the compressed version to suppress the influence of the compression (corresponding to the MOS correction introduced above).

The first step is to adapt the PSNR curve to the scale of MOS. To avoid infinite values of PSNR resulting from zero MSE if there was no error, we perform clipping of the PSNR to the value of 48dB, corresponding to MSE$=1$. According to our experience with subjective tests, people do not perceive the improvements resulting from higher PSNR. The PSNR curve needs to be smoothed and subsampled. Smoothing is necessary due to the human reaction time and the memory of the human eye; the PSNR curve contains a lot of sharp edges. To smooth the curve, we average over 1.5 sec fully overlapping intervals corresponding to the human reaction time. After the clipping and smoothing, the PSNR curve is normalized in the MOS scale by scaling factor a and shift factor b, that we obtained applying an affine minimum mean square error estimator. The optimal shift and scaling have then the form

$$a = \frac{c_{\text{MOS,PSNR}}}{\sigma^2_{\text{PSNR}}}, \qquad (4)$$

$$b = \mu_{\text{MOS}} - a \cdot \mu_{\text{PSNR}}. \qquad (5)$$

Here, $c_{\text{MOS,PSNR}}$ represents the sample covariance between the $\text{PSNR}[n]$ and the $\text{MOS}[n]$, μ_{PSNR} and μ_{MOS} are the sample means of PSNR respectively MOS and σ^2_{PSNR} is the sample variance of PSNR. We obtain the MOS estimation $\widehat{\text{MOS}}_{\text{PSNR}}[n]$ based on the PSNR:

$$\widehat{\text{MOS}}_{\text{PSNR}}[n] = a \cdot \text{PSNR}[n] + b. \qquad (6)$$

Note that the parameters a and b are proportional to the overall video sequence degradation (i.e. mean over $\text{PSNR}[n]$), which is not always proportional to the error probability (see Table I). To reliably approximate the dependency between a, b and the overall degradation, we would need larger data set. For the "news2" sequence, we obtained $a = 5.79$ and $b = 2$; the resulting $\widehat{\text{MOS}}_{\text{PSNR}}[n]$ can be seen in Figure 5. We can

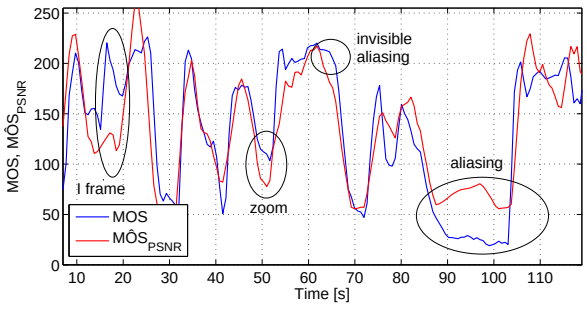

Fig. 5. Measured MOS and MOS estimated from PSNR ($\widehat{\text{MOS}}_{\text{PSNR}}[n]$).

see that $\widehat{\text{MOS}}_{\text{PSNR}}[n]$ already fits the MOS very well. The direct using of PSNR without smoothing, subsampling and scaling corresponding to the overall degradation resulted in correlations between 0.7 and 0.8. However, there are still some misalignments, some of them we marked by an ellipse within the Figure 5. Comparing the graphs with the videos for all test sequences, we can find the rules to explain the appearance of misalignments.

We measured the slope of the ascending and descending parts of the curves. The measured MOS curve is usually steeper when ascending than descending. This seems to be in contradiction with some previously published works stating that subjects are quicker to report a downward quality step, and slower to report an upward quality step [11]. However, in our case this result corresponds also to the reality – the downward quality steps in the video sequence are slow due to the error concealment. The user needs time to recognize an impairment from the real content. On the contrary, as soon

as the error free key frame arrives, it is a clearly visible and sudden improvement of the distorted stream; the change in quality is not gradual.

Especially critical are user to the effect we call "aliasing". It occurs if there is a packet loss during the rapid scene change (especially cut) and results in aliasing of two frames from different scenes. However, if there is only a slow local movement in the new scene, the user does not necessarily note that there should have been a scene change. This effect we call invisible aliasing (see Figure 6). If the impairment itself is almost as large as the whole screen, the correct part (movement) will be considered as an impairment. Invisible aliasing

Fig. 6. Screenshots from a part of the video stream with impairments caused by packet loss (top) and without them (bottom). A frame loss causes a degradation that results in low PSNR. However, the end-user recognizes such error much later.

results to a time shift between the measured and estimated MOS. Gradual scene changes like zooming or transition can also "conceal" some errors - it is not clear, what should be the part of the new scene and therefore end-user is not as critical to the error as is the PSNR based estimation. Table II summarizes the scenarios at which we systematically observed misalignments. Having observed these simple rules, we can

scenario	**condition**	**to correct**
sudden improvement	error free I frame, error free scene cut	ascending slope
invisible aliasing	low motion scene change	time
aliasing	packet loss in scene cut	value
transition/zooming	packet loss in zoom/transition	value

TABLE II

RULES FOR THE CORRECTION OF $\widehat{\text{MOS}}_{\text{PSNR}}[n]$, BASED ON THE HUMAN PERCEPTION.

correct the misalignments[1]. We perform following corrections:

1) In case of sudden improvement the slope of the ascending edge is corrected to s (if it is smaller). The slope is modified only within the 1.5 seconds long time interval, corresponding to the average measured time needed by the user to go up with the slider. The end of the so prolonged line is then connected to the nearest minima.
2) In case of invisible aliasing the shift n_s is added. The n_s seconds missing after the shift we replace by the last value.
3) Aliasing is penalized (without gradual changes) by multiplying all values with $k_a \in (0, 1)$.
4) Gradual scene changes are compensated by multiplying the minima with $k_g > 1$.

We obtained the parameters $s = 28.75$, $n_s = 2$s by averaging over our measured values for all sequences. We further obtained $k_a = 0.7$ and $k_g = 1.5$ by linear minimum mean square estimation applied to all tested sequences. The resulting predicted MOS ($\widehat{\text{MOS}}_{\text{corr}}[n]$) can be seen in Figure 7. Note

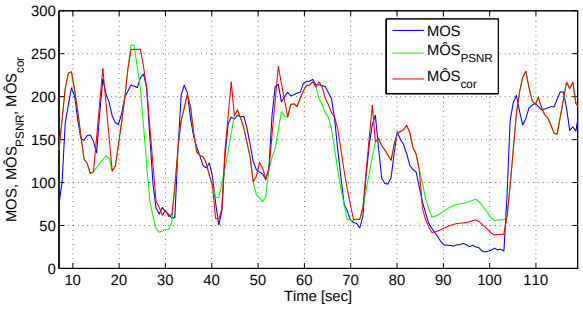

Fig. 7. Measured MOS, MOS estimated using modified PSNR ($\widehat{\text{MOS}}_{\text{PSNR}}[n]$) and MOS estimated using the corrections to $\widehat{\text{MOS}}_{\text{PSNR}}[n]$ based on the observed human vision rules ($\widehat{\text{MOS}}_{\text{corr}}[n]$).

that our corrections do not correct each misalignment, but in general (for different sequences), the similarity to the MOS curve is considerably increased.

V. EVALUATION OF RESULTS

To evaluate the performance of the video quality estimation, [2] recommends the linear (Pearson) correlation R_{lin} to express the prediction accuracy, the rank-order (Spearman) correlation R_{ro} to represent the monotonicity, and outlier ratio $R_{2\sigma} = N_{\text{out}}/N$ to check the consistency.

The length of the data vector is N, N_{out} is the number of outliers for which $\left|\text{MOS}[i] - \widehat{\text{MOS}}[i]\right| > 2\sigma_{\text{MOS}}$, where σ^2_{MOS} is the variance of the measured MOS. The evaluation of the proposed metric for different videos, compared to the scaled and smoothed PSNR $\widehat{\text{MOS}}_{\text{PSNR}}[n]$ for all the video sequences can be found in Table III. The effect of the corrections to the

Video sequence	R_{lin}	R_{ro}	$R_{2\sigma}$
"news 1"	0.9012	0.8752	0.0 %
"news 2"	0.9289	0.9105	0.0 %
"soccer"	0.8689	0.8563	0.1 %
"cartoon"	0.9304	0.9167	0.0 %
"city"	0.8644	0.8525	0.5 %
"traffic"	0.8311	0.8281	0.8 %
All sequences	0.8869	0.8644	0.0 %
$\widehat{\text{MOS}}_{\text{PSNR}}[n]$	0.8356	0.8307	0.2%

TABLE III

GOODNESS OF FIT FOR THE PROPOSED METRIC.

$\widehat{\text{MOS}}_{\text{PSNR}}[n]$ for the "news2" video sequence can be seen on

[1]The scene change detection and classification (cut/zoom/transition) is out of the scope of this article. Several methods can be found in [9], or in other literature dedicated to that topic.

the scatter plot in Figure 8. Scatter plots before and after the

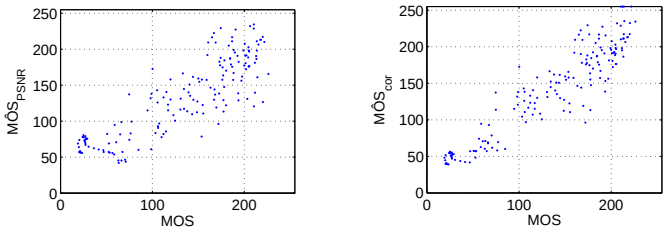

Fig. 8. Scatter plot for the measured MOS with $\widehat{\text{MOS}}_{\text{PSNR}}[n]$ (left) and $\widehat{\text{MOS}}_{\text{cor}}[n]$ (right) for the video sequence "news2".

correction for the whole sequence are visualized in Figure 9. It can be seen that the simple corrections are able to compensate the biggest mismatches between the perceived quality and the quality estimated by modified PSNR (scaled, clipped and smoothed). The $\widehat{\text{MOS}}_{\text{PSNR}}[n]$ already performs quite well.

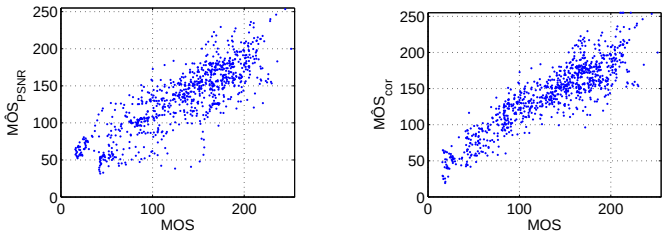

Fig. 9. Scatter plot for the measured MOS with $\widehat{\text{MOS}}_{\text{PSNR}}[n]$ (left) and $\widehat{\text{MOS}}_{\text{cor}}[n]$ (right) for all sequences.

VI. Conclusion

In this work we present a PSNR-based estimator of the time variant video quality for the low rate and low resolution video sequences, displayed on the mobile terminal. To obtain the mean opinion score, we performed subjective perceptual quality tests on UMTS mobile terminals using several video sequences with various data rates and packet loss probabilities. For these tests and obtaining MOS out of the user evaluations, we used a methodology adapted to the testing on mobile terminals, different to the methods recommended for the tests on PC screens. Using the proposed methodology, we were also able to obtain a consistent set of data. We mapped the resulting MOS curve on the appropriately scaled and smoothed PSNR and analyzed the differences. PSNR already provided 80-85% correlation with the measured data. Several differences turned out to be caused by the simple human perception rules.

After summarizing them, we proposed a new reference-based video quality metric – the PSNR corrected using simple rule-based algorithm. We evaluated the performance of our estimator by checking its accuracy, monotonicity and consistency for the six video sequences. The results show a considerable improvement compared to the PSNR. It would be beneficial to test our method with a data set containing more video sequences. However, obtaining such a set is rather time demanding and costly task. It is not possible to make long sessions with the test persons as their concentration decreases and the test results become higher variance. As further work we intent to perform tests including new data rate and error rate combinations, other video sequences and to investigate the dependency between the error rate and the appropriate PSNR scaling.

Acknowledgements

We thank mobilkom austria AG&CoKG for technical and financial support of this work. The views expressed in this paper are those of the authors and do not necessarily reflect the views within mobilkom austria AG&CoKG.

References

[1] ITU-R, "Recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures," September 1998.

[2] Video Quality Experts Group (VQEG), "Final Report from The VQEG on the Validation of Objective Models of Video Quality Assessment", March 2000, available in `http://www.vqeg.org/`.

[3] O. Nemethova , M. Ries, E. Siffel, M. Rupp, "Quality Assessment for H.264 Coded Low-Rate and low-Resolution Video Sequences," Proc. of Conf. on Internet and Inf. Technologies (CIIT), St. Thomas, US Virgin Islands, pp. 136-140, 2004.

[4] A.J. Seyler, Z.L. Budrikis, "Detail perception after scene changes in television image presentations," IEEE Trans. on Information Theory, Vol. 11, No. 1, pp. 31–42, Jan. 1965.

[5] D. Pearson, "Viewer response to time-varying video quality," Proc. of SPIE Human and Electronic Imaging III, vol. 3299, Apr. 1998.

[6] ANSI T1.801.03, "American National Standard for Telecommunications: Digital Transport of One-Way Video Signals. Parameters for Objective Performance Assessment," 2003.

[7] S. Winkler "Digital Video Quality: vision models and metrics", John Wiley & Sons Ltd., The Atrium, Chichester, England, 2005.

[8] M.H. Pinson, S. Wolf, "Comparing subjective video quality testing methodologies," Proceedings of SPIE Video Communications and Image Processing, Lugano, Switzerland, July, 2003.

[9] A. Hanjalic, "Content-Based Analysis of Digital Video," Kluwer Academic Publishers, 2004.

[10] O. Nemethova, M. Ries, A. Dantcheva, S. Fikar, M. Rupp, "Test Equipment for Time-Variant Subjective Perceptual Video Quality Testing with Mobile Terminals," in Proc. of International Conference on Human Computer Interaction (HCI 2005), Phoenix, USA, November, 2005.

[11] L. Gros, N. Chateau, "Instantaneous and Overall Judgements for Time-Varying Speech Quality: Assessments and Relationships," Acta Acustica, vol. 87, pp. 367-377, 2001.