NO-REFERENCE VIDEO QUALITY EVALUATION FOR HIGH-DEFINITION VIDEO

Christian Keimel, Tobias Oelbaum and Klaus Diepold

Technische Universität München, Institute for Data Processing, Arcisstr. 21, 80333, Munich, Germany
christian.keimel@tum.de, tobias.oelbaum@mytum.de, kldi@tum.de

ABSTRACT

A no-reference video quality metric for High-Definition video is introduced. This metric evaluates a set of simple features such as blocking or blurring, and combines those features into one parameter representing visual quality. While only comparably few base feature measurements are used, additional parameters are gained by evaluating changes for these measurements over time and using additional temporal pooling methods. To take into account the different characteristics of different video sequences, the gained quality value is corrected using a low quality version of the received video. The metric is verified using data from accurate subjective tests, and special care was taken to separate data used for calibration and verification. The proposed no-reference quality metric delivers a prediction accuracy of 0.86 when compared to subjective tests, and significantly outperforms PSNR as a quality predictor.

Index Terms— Visual quality metric, no-reference, AVC/H.264, SVC, HDTV

1. INTRODUCTION

Humans are able to judge the visual quality of a processed and distorted video without ever seeing the reference video. But subjective testing is time consuming, expensive, and can not be part of most practical applications. While full-reference (FR) and also reduced-reference (RR) video quality metrics are available that provide useful results, many systems or applications can not access the reference video and therefore require a no-reference (NR) evaluation. Additionally, the reference video may not even exist, as in the case of synthetic views generated for user selectable viewpoint applications. Furthermore, metrics using reference video sequences can not deliver accurate results, if the visual quality of the reference is not known. NR video quality evaluation has been the goal of many contributions in the field of visual quality metrics, but so far only limited results have been achieved. Major drawbacks of the presented approaches up to now are: only very few verification results that do not allow reasonable conclusions (e.g. [1]) , the use of the same data for design and verification of the metric (e.g. [2]) or the use of bit rate as quality indicator without using different encoders or at least different encoder settings (e.g. [3]). One popular approach for NR quality evaluation is the use of watermarks (e.g. [4]). Watermarks, however, need access to the reference video, and therefore can not be classified to be true NR metrics. Methods to predict the PSNR from the coded bit-stream (e.g. [5]) have been shown to work very well for High-Definition (HD) video, but are limited to the prediction accuracy of PSNR as a visual quality estimator.

In this contribution we present a no-reference video quality metric for HD video, encoded with AVC/H.264 and, partly, with its extension SVC. The metric can be split into two consecutive parts: in the first stage, we extract a set of features from the video, analyze the statistical distribution of these features to generate a set of parameters, and combine the parameters to a quality value. In the second stage, we correct this quality prediction using an additional video, created by encoding the received video to a low visual quality. The idea behind this correction step is, that while we do not have access to the original video, we can still generate a video with known quality. By comparing the received video to this low quality version, we are able to deduce the quality of the received video.

Our NR metric is verified using a set of seven different HD video sequences at a resolution of 1920×1080 pixel. These sequences were encoded using substantially different encoding settings to avoid tailoring the method to a special set of encoder options resulting in video exhibiting a wide range of different artifacts. The sequences were evaluated in carefully designed subjective tests, using a high number of observers and a controlled environment. In order to avoid the verification of the metric by using the same data used to build the model, we applied a cross validation approach, also known as "leave one out".

Section 2 describes the feature extraction from the video and the combination of the gained feature parameters into a quality value, before we introduce the proposed correction step in section 3. The verification process is described in section 4, before the results and a conclusion are presented in section 5 and section 6, respectively.

2. A NO-REFERENCE METRIC FOR HD VIDEO

The set of features extracted is very similar to the one used for our RR metric in [6]. The features are: blockiness, bluriness, activity and predictability. The first three features are measurements for what takes place in every single frame of the video. Predictability describes what occurs between the single frames of a video, and is based on the assumption, that visual quality is perceived differently, if transitions between neighboring frames are smooth or if abrupt changes appear.

2.1. Feature extraction

The algorithm for *blur* measurement [7] determines the width of an edge and calculates the blur by assuming that blur is reflected by wide edges. As blur is something natural in a fast moving video (motion blur), this measurement is adjusted by a simple piecewise linear correction if the video contains high amount of fast motion.

The algorithm for determining the *blocking* [8] calculates the horizontal and vertical blockiness by applying a Fourier transform along each line or column. The unwanted blockiness can be easily detected by the location in the spectra. The measured spectrum is compared to a smoothed version of the spectrum. Blockiness should appear as peaks at distinct frequencies. For both blur and blocking it is sufficient to only take into account the luminance channel.

 Digital Object Identifier: 10.1109/ICASSP.2009.4959791

The *activity* is assesed by measuring the amount of details according to the BTFR metric [9]. The percentage of turning points along each line and each row are calculated and then averaged to obtain one single value. As the amount of details noticed by an observer decreases with increasing motion, the activity measurement is adjusted if high motion is detected in the video. For simplicity, this measurement is performed only on the luminance channel.

In order to determine the *predictability*, a predicted image is created using a simple motion compensation based on block matching [10]. The actual image and its prediction are then compared block by block. To avoid, that single pixels dominate the SAD measurement, both images are filtered using first a Gaussian blur filter and a median filtering afterward. The output is the percentage of blocks that are not noticeable different.

To reduce the computational complexity, all features are extracted only for a sub-region of the frames. This sub-region is defined by the center cut of 1280×720 pixel of the 1920×1080 pixel frames.

2.2. Pooling

Discussion with test subjects who rated the distorted video sequences revealed, that there are three major artifacts which determine the visual quality of the video: bluriness, blocking, and obvious change in visual quality between neighboring frames.

The flickering effect is partly captured by the feature *predictability*, but analysis has shown, that this measurement can not describe the whole effect. Therefore we also calculate the difference in blur and blocking between neighboring frames, as these measurements vary significantly if two frames have a considerably different visual quality. Hence we use six different measurements: *blur*, *blocking*, *activity*, *predictability*, *dblur* and *dblocking*.

For each of these we calculate the mean for each frame and additional values describing the statistical distribution of these measurements for the whole video: maximum, minimum, 0.9 and 0.1 percentiles. This extra temporal pooling is motivated by the fact, that strong artifacts affect visual quality more than simple averaging suggests. Also the distribution of the feature measurements is not described well enough using only the mean. After dropping features that did not show a significant variation for different video sequences, we gained 22 different parameters, describing the statistical properties of the video sequences.

2.3. Model building

We used methods provided by multivariate data analysis to examine the data, and build stable prediction models out of these parameters. This approach was first proposed by Miyahara [11]. In particular, we used the principal component analysis (PCA) to get a more compact representation of the video, and a partial least squares regression (PLSR) to find the relationship between the principal components and the visual quality. We have already shown in [6], that this leads to stable and useful prediction models.

Before building the prediction model, all extracted parameters are first centered around their mean, as the interesting information does not lie in the absolute values, but in the variation of the parameters across different video sequences. Also all parameter values are scaled to have a standard deviation of 1.

The visual quality prediction $\widehat{y}$ can then be calculated as

$$\widehat{y} = b_0 + \mathbf{p} \cdot \mathbf{b}. \tag{1}$$

$\mathbf{b}$ is the column vector of the single estimation weights b_m for each parameter p_m. b_0 is the model offset.

Table 1: Weights for selected parameters

Parameter	none	Seeking	ParkJoy	Umbrella[a]
$blur_{mean}$	-0.003	-0.036	-0.201	-0.056
$blocking_{mean}$	-0.112	-0.146	-0.065	-0.169
$activity_{mean}$	0.413	0.148	0.747	0.406
$dblur_{max}$	-0.491	-0.181	-0.455	-0.317
$dblocking_{max}$	-0.460	-0.281	-0.121	-0.431

[a] Video sequence left out during the model building step

2.4. Cross calibration and model stability

The accuracy of video quality metrics should be tested using previously unknown video sequences, not included in the model building step. Therefore we did not build only one single model, but seven different ones, one for each of the seven different sequences used. As the database with only seven video sequences is comparably small, some of the sequences do have a significant influence on the models. Excluding one sequence from the model building step results in a significant change of the estimation weights b_m. To gain a more stable model, a larger number of different original video sequences are needed.

Selected weights for different models are given in Table 1. The weights show the influence of one parameter on the model. Large absolute values show, that this parameter does have a high influence on the visual quality. Negative values show, that while this parameter increases, the visual quality decreases. This is obvious for the values for blur and blocking. The data in Table 1 also shows, that the weights for the mean blur value differ by a factor of more than 60, depending on which video was excluded from the calibration set during the model building process. While the variance is not huge for other parameters, there is still a difference by a factor of more than four.

Hence each model delivers good prediction results in two cases:

1. The quality of a video included in the model building step should be predicted.
2. The quality of a unknown video very similar to one of the video sequences used to build the model should be predicted.

While the first case does not occur in real world scenarios, the second case happens but is very unlikely, especially if the database is comparably small. But even if a bigger database were available, it could happen, that one video to be evaluated has significant different properties than any of the video sequences used during the model building process.

3. CORRECTION STEP

To handle such cases, we propose to introduce a correction step similar to the correction step used in our RR metric [6]. It consists of evaluating a high quality and low quality version of the video using the same quality metric used to rate the actual video of interest. With this information, the general relationship between visual quality and the metric's output for the actual video is then estimated.

3.1. Additional video sequences

Clearly, we do not have access to the original video or any version of the video with high visual quality. But we can produce a low

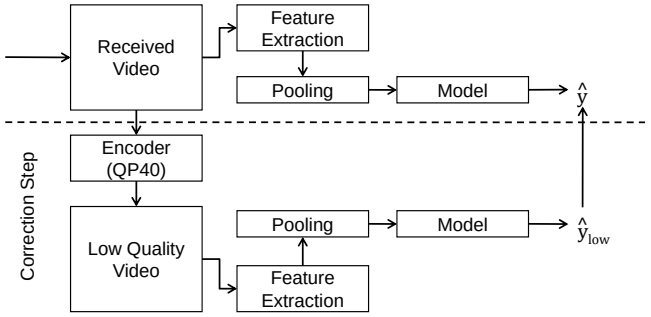

Fig. 1: No-reference prediction system

quality version of the received video. This version gives us some information about the "sensitivity" of this video to coding artifacts. Moreover, we know its expected quality, and by evaluating it we can perform a correction of the predicted quality. Hence we need for this correction the mean estimated quality value of the low quality version of the calibration video sequences $\overline{y_{low}}$, and the standard deviation σ_{low}. To avoid overcompensation, the quality prediction for this video, $\widehat{y}_{low}$, is first clipped to $\overline{y_{low}} \pm 3\sigma_{low}$ and the correction is then applied as follows:

$$\widehat{y} = \widehat{y} - (\widehat{y}_{low} - \overline{y_{low}}) * 0.75. \quad (2)$$

The factor of 0.75 was added to give more weight to the original prediction compared to the correction step introduced by the low quality instance of the video and determined experimentally. The overall prediction system is shown in Fig. 1. The low quality video was generated using a very simple fixed QP encoder for AVC/H.264. Using a very high QP of 40 ensures that the visual quality of this video is reasonably low. Finally, we slightly correct the prediction values $\widehat{y}$ using a fixed sigmoid nonlinear correction with $a = 1.0$, $b = 0.5$, $c = 0.2$, as for very good or very bad quality, subjective testing does have a nonlinear quality rating, and thus ratings do not reach the boundaries of the scale, but are saturated before. The general sigmoid function is given as

$$\widehat{y} = a/(1 + e^{(-(\widehat{y}-b)/c)}) \quad (3)$$

This correction function is nearly linear over a wide quality range and is not adapted to the actual data, but is a fixed part of the quality metric.

3.2. Effect of the correction step

The effect of the correction step is explained best on the video sequence "Seeking". This video shows blur values that are significantly above the values that were detected for all other sequences. Especially the maximum blur values are out of range: the mean value of $blur_{max}$ for all other sequences is around 3.4, whereas for "Seeking" this value is above 20. As a result of these "out of range" values, the predicted visual quality for the sequences (excluding "Seeking") is negative, whereas the visual quality normally should be in a range from 0 to 1. Calculating the visual quality for the low quality versions of these video sequences reveals that the quality was grossly underestimated: whereas $\overline{y_{low}}$ was found to be very close to zero, $y_{low,Seeking}$ was between -1.5 and -2.1. Using the above described correction step, the quality prediction $\widehat{y}$ was within the desired range of 0 to 1.

This correction step also significantly increases the prediction accuracy. It is interesting, that this correction is not necessary, if no cross validation is applied, and thus the same data is used for calibration and verification. Hence, the information given by the low quality instance enables the quality prediction of previously unknown video. This is also reflected by the low prediction accuracy achieved if the correction step is omitted (Table 2).

4. VERIFICATION

Verification of the proposed metric was done using seven different original HD video sequences encoded using the AVC/H.264 and SVC reference encoder. Significantly different encoder settings were applied to avoid training the models to one special encoder setting and also to account for a large quality range, resulting in bit rates from 4.5 Mbit/s up to 30 Mbit/s representing a quality range from "not acceptable" to "perfect" (0.21 to 0.94 on a 0 to 1 scale) and a total of 44 data points. From the SVT test set, the sequences CrowdRun, IntoTree, OldTownCross, ParkJoy, Seeking and Umbrella were used and additionally the sequence AlohaWave. All video sequences have a spatial resolution of 1920×1080 pixel and a temporal resolution of 25 or 50 frames per second (fps).

The subjective tests were performed at the video quality evaluation laboratory of the Institute for Data Processing at the Technische Universität München according to ITU-R BT.500 [12]. Some were performed as part of the official verification tests of SVC [13]. The tests were carried out using a variation of the standard DSCQS test method as proposed in [14]. The 95% confidence intervals of the subjective votes are below 0.07 on a 0 to 1 scale for all single test cases, the mean 95% confidence interval is 0.04.

5. RESULTS

The performance of the proposed no-reference metric is compared to PSNR and two FR metrics. The FR metrics are Edge-PSNR [9] and the video quality metric (VQM) according to Annex D of ITU-T J.144 [9]. For the VQM the general model and, due to limitations of the available VQM reference implementation, a progressive framerate of 25 fps was used for all video sequences.

In order to compare the prediction performance of different approaches the Pearson correlation, the Spearman rank order correlation and the root mean squared error (RMSE) was determined. For calculating the RMSE first order fitting was applied for the comparison metrics, while no fitting was used for the proposed NR method. A cross validation approach was used, ensuring that the results for every video are generated using a model calibrated with a database that does not contain this particular video. For comparison, results for the model if no cross validation had been applied are also given.

The results in Table 2 show, that the proposed NR metric (Fig. 2) significantly outperforms PSNR. The high prediction accuracy of 0.86 is not a result of fitting the model to the actual data, but shows the real prediction accuracy, as a cross validation approach was used to gain the performance numbers. The importance of using such a cross validation approach is shown by the unrealistically high prediction accuracy of 0.95 if this step is omitted. The effectiveness of the correction step as proposed in section 3, is shown by the fact, that without this correction, the quality for some sequences is grossly over- or underestimated, and the correlation to subjective results is rather low.

The rather bad performance of the ITU-T J.144, Annex D VQM compared to PSNR might be explained by its limitation to 25 fps, considering we also included material with originally 50 fps.

Table 2: Prediction results

	Pearson	Spearman	RMSE[(a)]
NR no cross validation	0.95	0.94	0.06
NR no correction step	0.51	0.66	0.35
Proposed NR	0.86	0.85	0.11
PSNR	0.69	0.69	0.14
ITU-T J.144, Annex D	0.62	0.69	0.15
Edge-PSNR	0.70	0.70	0.13

[(a)] After first order fitting for all comparison metrics, no fitting for the proposed NR metric

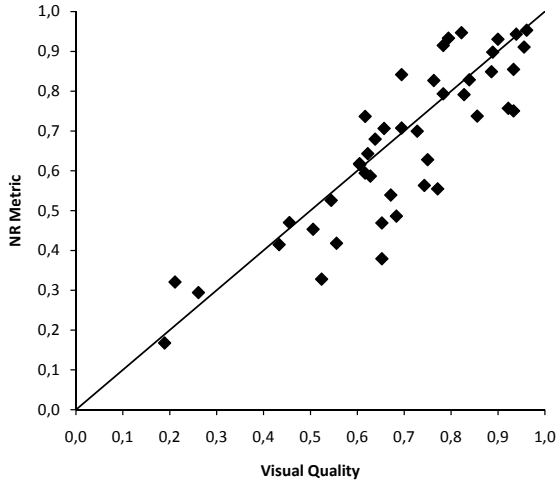

Fig. 2: Proposed NR metric

6. CONCLUSION

We proposed a new no-reference quality metric for High-Definition video. The metric is based on a set of simple features and respective parameters that are extracted from the video and a subsequent correction step. This correction step uses a low quality video that is produced by encoding the received video. The metric was verified on a set of HD video sequences that were encoded with AVC/H.264 or SVC. The visual quality of these video sequences was determined in precise subjective tests and a cross validation approach was used to separate the data used for calibration and verification. Results show, that the proposed no-reference metric significantly outperforms PSNR, and performs equally well as the best full-reference comparison metric.

Selection of the features and parameters that are used for the no-reference metric is not explicitly based on a Human Visual System (HVS) model, but we selected the features and parameters that are best suited to describe the variation of the video sequences and the quality variation. Also the model building step does not follow a HVS based approach, but is based on data modeling methods such as the PCA and the PLSR. The presented results show the effectiveness of this approach.

Whereas the first part of the metric does not contain any substantially new aspects, and is very similar to existing no-reference metrics [1, 2], the presented approach differs in two main aspects from these works. Firstly, the introduced correction step allows to predict the quality of previously unknown sequences, and significantly increases the prediction accuracy. Secondly, the data used to calibrate the metric was separated from the data used for verification, thus resulting in a meaningful verification of the metric.

7. REFERENCES

[1] M. Montenovo, A. Perot, M. Carli, P. Cicchetti, and A. Neri, "Objective quality evaluation of video services," in *Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2006.

[2] S. Péchard, D. Barba, and P. Le Callet, "Video quality model based on a spatio-temporal features extractions for H.264-coded HDTV sequences," in *Proc. Picture Coding Symposium*, Nov. 2007.

[3] M. Ries, O. Nemethova, and M. Rupp, "Performance evaluation of video quality estimators," in *Proc. European Signal Processing Conference EUSIPCO*, Sep. 2007, pp. 159–163.

[4] Y. Fu-Zheng, W. Xin-Dai, C. Yi-Lin, and W. Shuai, "A no-reference video quality assessment method based on digital watermark," in *Proc. 14th IEEE Personal, Indoor and Mobile Radio Communications 2003*, vol. 3, Sep. 2003, pp. 2707–2710.

[5] A. Eden, "No-reference estimation of the coding PSNR for H.264-coded sequences," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 667–674, May 2007.

[6] T. Oelbaum and K. Diepold, "A reduced reference video quality metric for AVC/H.264," in *Proc. European Signal Processing Conference EUSIPCO*, Sep. 2007, pp. 1265–1269.

[7] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 3, Sep. 2002, pp. 57–60.

[8] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 3, Oct. 2000, pp. 981–984.

[9] *ITU-T J.144. Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*, ITU-T Std., Mar. 2004.

[10] T. Zahariadis and D. Kalivas, "Fast algorithms for the estimation of block motion vectors," in *Proceedings of the Third IEEE International Conference on Electronics, Circuits, and Systems, ICECS*, Oct. 1996, pp. 716–719.

[11] M. Miyahara, "Quality assessments for visual service," *IEEE Communications Magazine*, vol. 26, no. 10, pp. 51–60, 1988.

[12] *ITU-R BT.500 Methodology for the Subjective Assessment of the Quality for Television Pictures*, ITU-R Std., Rev. 11, Jun. 2002.

[13] MPEG Test Subgroup, "SVC verification test report," ISO/IEC JTC1/SC29/WG11, Tech. Rep. N9577, Jan. 2008. [Online]. Available: {http://www.chiariglione.org/mpeg/quality_tests.htm}

[14] V. Baroncini, "New tendencies in subjective video quality evaluation," *IEICE Transaction Fundamentals*, vol. E89-A, no. 11, pp. 2933–2937, Nov. 2006.