LENGTH-INDEPENDENT REFINEMENT OF VIDEO QUALITY METRICS BASED ON MULTIWAY DATA ANALYSIS

Clemens Horch, Christian Keimel, Julian Habigt and Klaus Diepold

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

ABSTRACT

In previous publications it has been shown that no-reference video quality metrics based on a data analysis approach rather than on modeling the human visual system lead to very promising results and outperform many well-known full-reference metrics. Furthermore, the results improve when taking the temporal structure of the video sequence into account by using multiway analysis methods. This contribution shows a way of refining these multiway quality metrics in order to make them more suitable for real-life applications and maintaining the performance at the same time. Additionally, our results confirm the validity of H.264/AVC bitstream no-reference quality metrics using multiway PLSR by evaluating this concept on an additional dataset.

Index Terms— Video quality metric, no-reference metric, multiway data analysis, multiway PLSR, trilinear PLS.

1. INTRODUCTION

The traditional method for designing video quality metrics tries to model the human visual system (HVS) in order to reproduce the perception of a human observer. This requires a sufficient understanding of the HVS, which we currently do not possess. Therefore, we choose the data-driven approach, which does not require this knowledge. There have already been some contributions using a data analysis approach to build a no-reference video quality metric. Our most promising metrics are based on the extraction of H.264/AVC bitstream features from the video data which are used to train a regression model [1, 2, 3]. Recently we have shown that the inclusion of the temporal dimension by using multiway data analysis further improves the prediction [1]. Mathematically, this can be achieved using extended regression methods like the two-way version of principal component regression (2D-PCR) [4] or multiway partial least squares regression (multiway PLSR) [2]. PLSR itself has already been used for the design of video quality metrics in [5, 6].

Although the metrics based on H.264/AVC bitstream features show very good performance, they also have some drawbacks. In general, the usage of H.264/AVC bitstream features limits the metric to video material that has been encoded with this technology. This is acceptable since H.264/AVC nowadays is the predominant video encoding standard. Furthermore, the inclusion of the temporal dimension into the prediction model requires all training sequences to consist of the same number of frames and also the sequence whose quality is to be predicted needs to match this length. This problem impedes the application of the metric in most fields as it is not feasible to train different regression models for each occurring length of video sequences.

In this contribution we show a way around this second problem by splitting all used video sequences into subsets of equal lengths. A group of pictures (GOP) seems to be an adequate choice for the length of those subsets. Consequently, the quality metric predicts a quality value for each GOP of the video sequence. A quality prediction for the complete sequence is calculated by taking the average of the per-GOP quality values. This can be seen as an analogy to subjective testing, where each observer gives a rating of the overall quality of the complete video sequence.

Features extracted directly from the H.264/AVC bitstream have been used by Eden [7] to estimate the PSNR of interlaced HDTV video or Slanina et al. [8], who estimate the PSNR of video sequences in CIF resolution. Rossholm and Lövström [9] used bitstream features to estimate some other quality metrics additional to the PSNR. Keimel et al. [1, 2, 3] refined the usage of bitstream features in order to directly estimate the visual video quality.

In the following, we will first give an overview of the feature extraction step, then discuss the multiway PLSR and point out the changes that are required to make the metric length-independent. Finally, we will show the validity of the metric by comparing its prediction performance to other no-reference metrics.

2. DESIGN OF THE VIDEO QUALITY METRIC

A video quality metric that is based on a data analysis approach uses feature-data as input to train a prediction model. After the training step the model can be used to predict the quality of unknown video sequences from their features. The feature data used during the training is represented by the $N \times M \times T$ array $\underline{\mathbf{X}}$, where N denotes the number of sequences, M the number of features and T the length of the video sequences in frames. The $N \times 1$ column vector $\mathbf{y}$ contains the subjective quality determined in subjective tests as ground truth. The aim is to find the unknown $M \times 1 \times T$ weight array $\underline{\mathbf{B}}$ that expresses the visual quality using only the feature data:

$$\mathbf{y} = \frac{1}{T} \sum_{t=0}^{T} \underline{\mathbf{X}}(:,:,t)\underline{\mathbf{B}}(:,:,t) \tag{1}$$

2.1. H.264/AVC Bitstream Feature Extraction

A modified version of the H.264/AVC JM reference software is used to extract $M = 17$ different features per slice resulting in an $M \times T$ feature matrix per sequence. The feature extractor parses the Network Abstraction Layer (NAL) of the H.264/AVC byte stream and extracts some statistical data from the Video Coding Layer (VCL) after reversing the entropy coding. Over the course of the extraction, it descends from the slice layer to the macroblock and submacroblock

 Digital Object Identifier (DOI): 10.1109/ICIP.2013.6738010

layers. The following list gives a short overview of the extracted features, a more detailed explanation can be found in [2, 3]:

- Slice type: I-, P- or B-slice
- Size of the slice in kilobits
- Average QP per slice and variation of QP in the slice
- Percentages of different macro- and submacroblock types per slice
- Motion vector lengths (average and maximum)
- Motion vector errors (average and maximum)

In this contribution we focus on coding artifacts rather than quality degradation caused by the transmission of video. The reason for this is that our feature extraction tools only works reliably on intact bitstreams.

2.2. Trilinear Partial Least Squares Regression

In the case of a two-way feature matrix, principal component regression (PCR) or bilinear partial least squares regression (PLS1) are suitable methods to build a regression matrix. In the three-way case we focus on, trilinear partial least squared regression (Tri-PLS1) is required. This multidimensional extension of PLS1 was introduced by Bro [10]. In Tri-PLS1 the components are determined depending on weights gained along both the m and the t dimension, whereas in PLS1 the components are only dependent on the m dimension.

Algorithm 1 shows an iterative algorithm that describes the decomposition of $\underline{\mathbf{X}}$ into its components $\mathbf{w}^M$ and $\mathbf{w}^T$ along both feature dimensions. $\mathbf{Z}$ in step 2 of the algorithm represents the matrix of all z_{mt} with

$$z_{mt} = \sum_{n=1}^{N} y_n x_{nmt}. \tag{2}$$

The scores t_n corresponding to each sample n can then be written with the components as

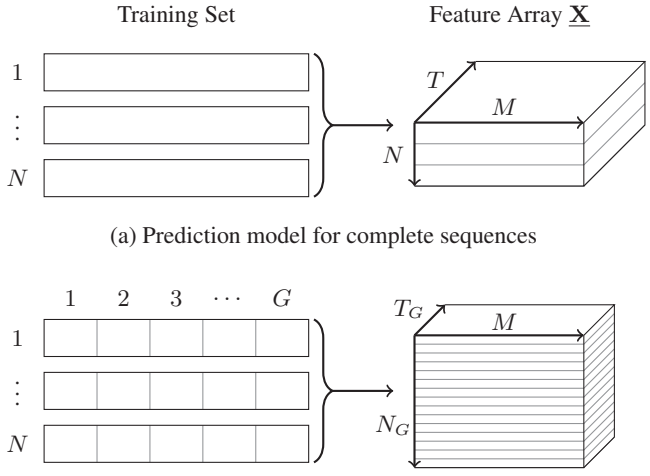
$$t_n = \sum_{m=1}^{M} \sum_{t=1}^{T} x_{nmt} w_m^M w_t^T. \tag{3}$$

Algorithm 1 Trilinear PLSR (Tri-PLS1)

center $\underline{\mathbf{X}}$ and $\mathbf{y}$
$\underline{\mathbf{X}}_1 = \underline{\mathbf{X}}, \mathbf{y}_1 = \mathbf{y}$
$f = 1$
1: **repeat**
2: calculate $\mathbf{Z}$
3: determine $\mathbf{w}_f^m$ $\mathbf{w}_f^t$ by SVD of $\mathbf{Z}$
4: calculate $\mathbf{t}_f$. $\mathbf{T} = [\mathbf{t}_1 \cdots \mathbf{t}_f]$
5: $\mathbf{b}_f = (\mathbf{T}^\top \mathbf{T})^{-1} \mathbf{T} \mathbf{y}_f$
6: $\underline{\mathbf{X}}_{f+1} = \underline{\mathbf{X}}_f - t_f \mathbf{w}_f^m (\mathbf{w}_f^t)^\top$ and $\mathbf{y}_{f+1} = \mathbf{y}_f - \mathbf{T}\mathbf{b}_f$
7: $f = f + 1$
8: **until** proper description of $\mathbf{y}_f$

From the extracted components and scores, we can then obtain an estimate of the $T \times M$ weight matrix $\hat{\mathbf{B}}$ for direct regression of an $1 \times M \times T$ slice of the feature array $\mathbf{X}_u$ representing the features of an unknown sequence. The quality estimation $\hat{y}_u$ for this unknown sequence can be written as

$$\hat{y}_u = \hat{b}_0 + \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{u,t} \hat{\mathbf{b}}_t \tag{4}$$



(a) Prediction model for complete sequences

(b) Prediction model based on GOP subsets

Fig. 1. Length-independent quality prediction

where $\mathbf{x}_{u,t}$ denotes the t-th column vector of $\mathbf{X}_u$ and $\hat{\mathbf{b}}_t$ the t-th row vector of $\hat{\mathbf{B}}$

2.3. Length-Independent Quality Estimation

The matrix multiplication in (4) requires that $\mathbf{X}_u$ and $\hat{\mathbf{B}}$ share the same dimension T. This means that the video sequences used in the training of the regression model need to have the exact same number of frames as the unknown sequence. For any real application of the quality metric this is a severe drawback, since this condition generally cannot be fulfilled.

To solve this issue, we propose to split up the video sequence in small subsets of equal length and predict the quality of each subset individually. In general it is advisable to split H.264/AVC-encoded video only at intra-coded frames (I-Frames) – a procedure that seems sensible in this case as well, especially as we use H.264/AVC bitstream features.

In our improved prediction model we split all video sequences into G subsets of T_G frames, the length of one GOP. As a consequence, we obtain an $N_G \times M \times T_G$ feature array $\underline{\mathbf{X}}$ to train the model, where N_G denotes the total number of video subsets. If all sequences in the training set have the same length, then $N_G = NG$. Fig. 1 illustrates the difference in dimensionality between the feature arrays depending on whether complete sequences or GOP-subsets are used. In summary, the training set consists of much more but also much shorter sequences.

Since we do not have subjective quality data per GOP, the same value y_n is used for all subsets that were cut from the video sequence with the MOS value y_n. The vector of the dependent variables $\mathbf{y}$ is expanded accordingly from $N \times 1$ to $N_G \times 1$ by duplicating elements.

The training of the regression model itself remains unchanged and is done as described in section 2.2. To predict the quality $\hat{y}_u$ of an unknown sequence, the sequence is split up into its G individual GOPs of length T_G. Clearly T_G has to be the same as in the training but this is much more likely than assuming the equal number of frames for the complete sequence. After predicting the quality for each GOP, $\hat{y}_u$ is set to the mean of the predictions of all GOP-subsets. Fig. 2 shows an example for the prediction results of the proposed metric. In particular, it illustrates that the average of

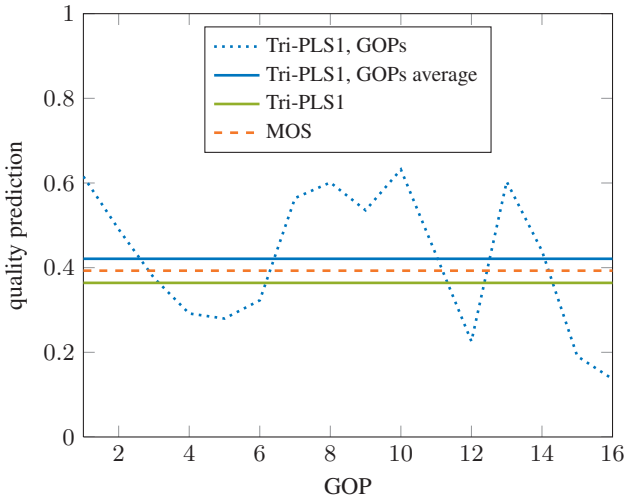


Fig. 2. Example for the different prediction models, video sequence *Stephan* at 265 kbit/s

the per-GOP quality prediction is fairly close to the quality estimation of a Tri-PLS1 metric applied on the complete sentence and the subjective MOS value.

2.4. Sigmoid Correction

In subjective testing, the ratings at the boundaries of the rating scale show a nonlinear nature and saturate much earlier. To take this into account, we correct all prediction values $\hat{y}$ using a fixed sigmoid correction function [5]

$$\hat{y}_S = \frac{1}{1 + e^{-5 \cdot (\hat{y} - 0.5)}}. \tag{5}$$

This function is not adapted to the actual data, but is rather a fixed part of the quality metric. $\hat{y}_S$ represents the final result of the quality prediction.

3. PERFORMANCE EVALUATION

In order to demonstrate and evaluate the performance of the proposed metric, we compare it to other data analysis based metrics and some other well-known video quality metrics. All metrics were applied to the same dataset described in the following.

3.1. Dataset Used for Evaluation

For the evaluation of the proposed metric, we used parts of the dataset provided by IT-IST [11]. We chose 4 different bitrates in the range from 32 kbit/s to 2048 kbit/s for each of the 12 different video sequences in the dataset: Australia, City, Coastguard, Container, Crew, Football, Foreman, Mobile, Silent, Stephan, Table and Tempete (cf. Fig. 4). Hence the set of videos consisted of $N = 48$ video sequences in common intermediate (CIF) resolution (352×288). The sequences had been encoded using H.264/AVC with a fixed GOP-length of $T_G = 15$ frames. We removed the first GOP, as it consists of only 13 frames, and also the last few frames because the last GOP is incomplete. In total there are $T = 240$ frames per sequence and $G = 16$ GOPs per sequence which gives $N_G = 768$ subsets of video to train the model with.

IT-IST provides the results of subjective quality assessment for all video sequences in their data set. The test was conducted with 42 participants using degradation category rating as described in ITU-T Recommendation P.910. We assume that the MOS values are equally valid for the slightly shortened sequences we used.

To evaluate the performance of the new metric (referred to as *Tri-PLS1-GOP*), the multi-way quality prediction without splitting the sequences into GOPs as described in [2] (*Tri-PLS1)* was also applied to the data set. For further comparison the well-known full-reference video quality metric SSIM [12] and the PSNR were calculated for the data set as well.

3.2. Cross Validation

In the evaluation of data analysis methods it is important not to use the same data for the training and the validation. Therefore, we performed cross validation by leaving out 4 video sequences of the same content and used the remaining 42 as the training set. Afterwards, the quality of the sequences that had been left out were predicted using this model. In doing so, we obtain quality predictions for all 48 sequences without using the same data for training and validation.

3.3. Results

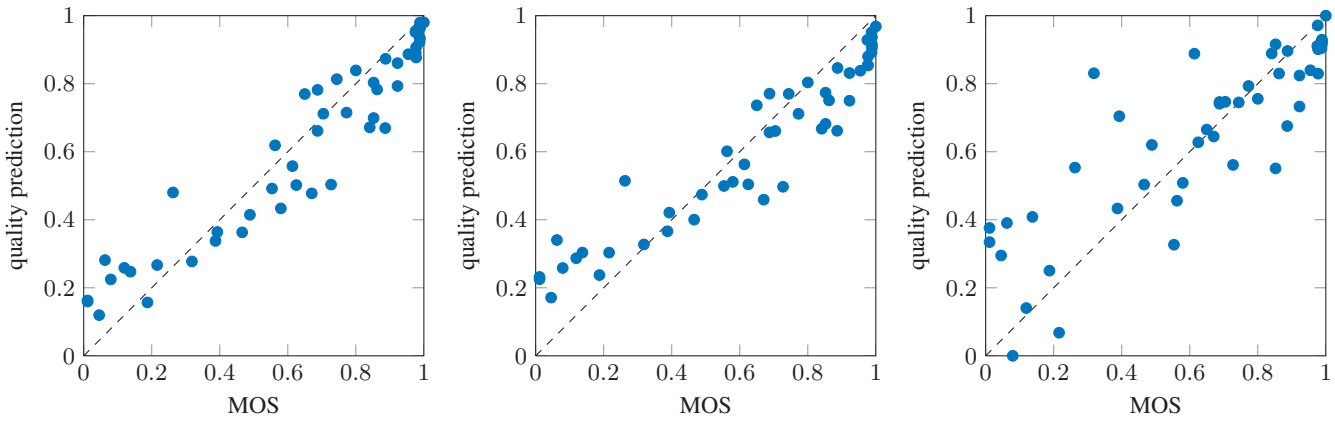
The performance of the quality metrics is evaluated by calculating both the Pearson correlation coefficient and the Spearman rank order correlation coefficient between the predicted quality $\hat{\mathbf{y}}$ and the corresponding subjective results $\mathbf{y}$. The values for the proposed metrics are shown in Table 1 along with the Root Mean Square Errors (RMSE). Fig. 3 shows scatter plots of the quality estimates against the mean opinion scores for Tri-PLS1, Tri-PLS1-GOP and the full-reference metric SSIM.

Both the Pearson and the Spearman coefficients suggest a slight decrease in prediction performance when using Tri-PLS1-GOP instead of Tri-PLS1. The same statement can be made when looking at the RMSE. Nevertheless, the difference of the correlations is only visible in the third decimal digit and generally the correlation is on a very high level with both methods. Thus, one can hardly speak of a real disadvantage of the GOP-method – especially when we consider the advantage of length-independence. Moreover, the length-independence also allows us to estimate the visual quality per GOP, while still maintaining the same overall quality prediction as the Tri-PLS1 as shown in Fig. 2.

Table 1. Prediction performance

	Pearson	Spearman	RMSE
PSNR	0.723	0.777	0.346
SSIM [12]	0.850	0.871	0.172
Brandão et al. [11]	0.938	0.949	0.110
PLS1 [3]	0.935	0.919	0.117
tri-PLS1	0.951	0.962	0.108
tri-PLS1-GOP	0.947	0.955	0.125

Both metrics clearly outperform the full-reference metrics PSNR and SSIM. In comparison to the no-reference metric by Brandão and Queluz [11] based on the same dataset, our metric shows better correlation values for quality estimation on the used dataset.



(a) Tri-PLS1 (b) Tri-PLS1-GOP (c) SSIM

Fig. 3. Scatter plots of the compared metrics Tri-PLS1, Tri-PLS1-GOP and the full-reference metric SSIM

In addition, we performed quality estimation on the same dataset with a PLS1-based metric similar to the one described in [3]. The dimensionality of the three-way feature array was reduced by temporal pooling as discussed in [1]. On the one hand the results confirms the results from [3] – the PLS1 metric produces similar correlation values for the dataset used here. On the other hand this again shows that the addition of the temporal dimension leads to better results. In our case especially the Spearman coefficient increases clearly.

4. CONCLUSION

We improved the design of video quality metrics using multiway data analysis by making the training and the quality prediction independent of the length of video sequences. The proposed no-reference metric is based on features extracted from H.264/AVC bitstreams and makes use of the GOP-structure of H.264/AVC encoded video.

It turns out that averaging the per-GOP estimated quality values of a video sequence results in a quality estimation that correlates very well with the perceived quality as measured in subjective tests. Our results show that this metric performs equally well as a corresponding length-dependent metric and outperforms common full-reference metrics.

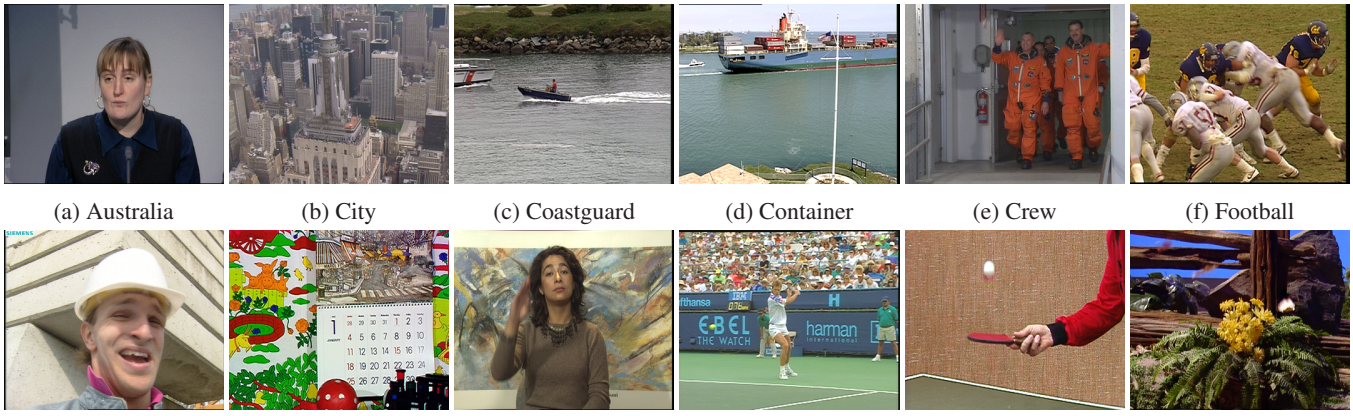
Apart from that, the main advantage of the presented metric is its improved universality when it comes to real world application. What remains is the drawback that all video sequences need to be encoded with the same GOP-length, but this is less inconvenient than demanding equal lengths for the complete sequences and very common in broadcasting applications.

Additionally, our results confirm the validity of the multiway PLSR-based metric previously presented by Keimel et al. [2].

5. REFERENCES

[1] C. Keimel, M. Rothbucher, Hao Shen, and K. Diepold, "Video is a cube," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 41–49, Sept. 2011.

[2] C. Keimel, J. Habigt, M. Klimpke, and K. Diepold, "Design of no-reference video quality metrics with multiway partial least squares regression," in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, Sept. 2011, pp. 49–54.



(a) Australia (b) City (c) Coastguard (d) Container (e) Crew (f) Football

(g) Foreman (h) Mobile (i) Silent (j) Stephan (k) Table (l) Tempete

Fig. 4. IT-IST Dataset

[3] C. Keimel, M. Klimpke, J. Habigt, and K. Diepold, "No-reference video quality metric for HDTV based on H.264/AVC bitstream features," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, Sept. 2011, pp. 3325–3328.

[4] C. Keimel, M. Rothbucher, and K. Diepold, "Extending video quality metrics to the temporal dimension with 2D-PCR," in *Image Quality and System Performance VIII*, Susan P. Farnand and Frans Gaykema, Eds. Jan. 2011, vol. 7867 of *Proceedings of SPIE*, SPIE.

[5] C. Keimel, T. Oelbaum, and K. Diepold, "No-reference video quality evaluation for high-definition video," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, Apr. 2009, pp. 1145–1148.

[6] C. Keimel, T. Oelbaum, and K. Diepold, "Improving the prediction accuracy of video quality metrics," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, Mar. 2010, pp. 2442–2445.

[7] A. Eden, "No-reference estimation of the coding PSNR for H.264-coded sequences," *Consumer Electronics, IEEE Transactions on*, vol. 53, no. 2, pp. 667–674, May 2007.

[8] M. Slanina, V. Ricny, and R. Forchheimer, "A novel metric for H.264/AVC no-reference quality assessment," in *Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on*, June 2007, pp. 114–117.

[9] A. Rossholm and B. Lövström, "A new video quality predictor based on decoder parameter extraction," in *International Conference on Signal Processing and Multimedia Applications*. Inst Syst & Technologies Informat, Control & Commun, 2008, pp. 285–290.

[10] R. Bro, "Multiway calibration. multilinear PLS," *Journal of Chemometrics*, vol. 10, no. 1, pp. 47–61, Oct. 1996.

[11] T. Brandão and M. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.

[12] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, Apr. 2004.