HYBRID NO-REFERENCE VIDEO QUALITY METRIC BASED ON MULTIWAY PLSR

Christian Keimel, Julian Habigt and Klaus Diepold

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

ABSTRACT

In real-life applications, no-reference metrics are more useful than full-reference metrics. To design such metrics, we apply data analysis methods to objectively measurable features and to data originating from subjective testing. Unfortunately, the information about temporal variation of quality is often lost due to the temporal pooling over all frames. Instead of using temporal pooling, we have recently designed a H.264/AVC bitstream no-reference video quality metric employing multiway Partial Least Squares Regression (PLSR), which leads to an improved prediction performance. In this contribution we will utilize multiway PLSR to design a hybrid metric that combines both bitstream-based features with pixel-based features. Our results show that the additional inclusion of the pixel-based features improves the quality prediction even further.

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

1. INTRODUCTION

The overall goal of video quality metrics is to obtain an accurate model of the spatial and temporal properties of the human visual system (HVS). This allows us to predict the quality as perceived by human observers adequately. But in order to do this, we assume that the HVS is understood well enough to create an adequate model. Considering a more data-driven approach, we regard the HVS as a black box: we use data analysis methods to determine the relationship between objectively measurable features at its input and the subjective quality at the box's output. One such data analysis method is Partial Least Squares Regression (PLSR).

Objective features are often determined on a frame-by-frame basis and then temporally pooled over all frames before applying the data analysis. This, however, neglects the temporal nature of video in the process. Pooling, especially averaging, obscures the influence of temporal distortions on the quality perception and thus leads to less than optimal models. The data analysis concept was hence expanded into the temporal dimension in [1] with multiway PLSR and its use in the design of a no-reference bitstream-based video quality metric, thus avoiding the temporal pooling step. Yet, by only using bitstream features, we ignore the information about visible distortions contained in the reconstructed frames themselves.

In this contribution, we will therefore extend the bitstream metric proposed in [1] with pixel-based features, leading to a hybrid no-reference video quality metric for HDTV. Of course this hybrid metric will be limited in its overall application as it is only designed to work with a specific coding technology, in this case H.264/AVC. But as this standard is the predominant coding technology for HDTV, the metric can still be considered to be fit for real-life applications.

In related works, Yamagishi et al. present a no-reference hybrid metric targeting IPTV in [2], but it uses only simple spatial and temporal activity for its pixel-based part. The same activity measurements are also used by Sugimoto et al. in [3] for a hybrid metric aimed at interlaced HDTV and thus more closely related to our contribution. In [4], Farias et al. also consider blockiness and bluriness in addition to bitstream features, similar to our contribution, but they only examine videos in CIF resolution. The development of no-reference hybrid video quality metrics is also a focus of ongoing research within the Video Quality Experts Group's (VQEG) Joint Effort Group (JEG) [5].

We will discuss in the first section PLSR and multiway PLSR, before introducing the bitstream and pixel-based feature extraction. This will be followed by a description of the design process of our hybrid metric. Before we conclude with a short summary, we present and discuss the results of the proposed metric.

2. DESIGN OF VIDEO QUALITY METRICS WITH PARTIAL LEAST SQUARES REGRESSION

In our data-driven approach, we do not assume a-priori specific relationships between the features and the visual quality, but rather gain the relationships by analyzing the available data. Firstly, we construct a data matrix $\mathbf{X}$ where the rows correspond to data from individual sequences and the columns represent the bitstream or pixel-based features. The visual quality values that were determined in subjective tests

are represented by the $n \times 1$ column vector $\mathbf{y}$. With n sequences and m features, $\mathbf{X}$ is an $n \times m$ matrix. Our aim is to find the unknown $m \times 1$ regression weight vector $\mathbf{b}$, mapping the features to the visual quality

$$\mathbf{y} = \mathbf{X}\mathbf{b}. \tag{1}$$

For more information about our data-driven approach, we refer to [6].

2.1. Bilinear Partial Least Squares Regression

PLSR is an extension of the principal component regression method (PCR). For PCR, the data matrix $\mathbf{X}$ is first subjected to a principal component analysis (PCA), and then for selected principal components (PC) a regression on $\mathbf{y}$ is done. The disadvantage of PCR is that the PCs best suited to represent $\mathbf{X}$, carrying the structure of the videos, are not necessarily the same PCs best suited to explain the variance in $\mathbf{y}$, describing the quality variation of the videos. In contrast, the modeling with PLSR is done simultaneously on $\mathbf{X}$ and $\mathbf{y}$, ensuring PCs that explain the variance in both $\mathbf{X}$ and $\mathbf{y}$ best. This basic type of PLSR is also called bilinear partial least squares or PLS1 . Based on the extracted PCs, we can then obtain an estimation $\hat{\mathbf{b}}$ of the regression weight vector $\mathbf{b}$ and thus can write the quality estimation $\hat{\mathbf{y}}$ for $\mathbf{y}$ as

$$\hat{\mathbf{y}} = \mathbf{1}\hat{b}_0 + \mathbf{X}\hat{\mathbf{b}} + \mathbf{e}, \tag{2}$$

where $\mathbf{1}$ describes the identity matrix, $\hat{b}_0$ the model offset and $\mathbf{e}$ the estimation error of the model. For unknown video sequences with a $1 \times m$ feature vector $\mathbf{x}_u$, the quality can then be predicted as

$$\hat{y}_u = \hat{b}_0 + \mathbf{x}_u\hat{\mathbf{b}}. \tag{3}$$

For more information on PLS1, we refer to [7].

2.2. Trilinear Partial Least Squares Regression

The multidimensional extension of PLS1, multiway or N-way PLS, was introduced by Bro in [8]. It extends the principle behind PLS1 of maximizing the variance explained by the PCs in both sides of (1) to higher dimensional data. In particular, the trilinear partial least squares (Tri-PLS1) describes the partial least squares regression of a three-way $n \times m \times t$ data array $\mathbf{X}(:,:,:)$ onto an $n \times 1$ column (quality) vector $\mathbf{y}$. The main difference compared to PLS1 is that the principal components are now determined dependent on weights gained along both the m and t dimension, whereas in PLS1 the principal components are only dependent on the m dimension.

The iterative algorithm shown in Listing 1 describes how $\mathbf{X}(:,:,:)$ is decomposed in its PCs $\mathbf{w}^m$ and $\mathbf{w}^t$ along both feature dimensions. $\mathbf{Z}$ represents the matrix of all z_{mt}, with

$$z_{mt} = \sum_{n=1}^{N} y_n x_{nmt}. \tag{4}$$

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

$$t_n = \sum_{m=1}^{M}\sum_{t=1}^{T} x_{nmt} w_m^M w_t^T. \tag{5}$$

Algorithm 1: Trilinear PLS1

center $\mathbf{X}$ and $\mathbf{y}$
$\mathbf{y}_0 = \mathbf{y}$
$f = 1$

1 Calculate $\mathbf{Z}$
2 Determine $\mathbf{w}_f^m$ and $\mathbf{w}_f^t$ by SVD of Z
3 Calculate $\mathbf{t}_f$. $\mathbf{T} = [\mathbf{t}_1 \cdots \mathbf{t}_f]$
4 $\mathbf{b}_f = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}\mathbf{y}_0$
5 Each sample $\mathbf{X}_i$ is replaced with $\mathbf{X}_i - t_i\mathbf{w}_f^m(\mathbf{w}_f^t)^T$ and $\mathbf{y} = \mathbf{y}_0 - \mathbf{T}\mathbf{b_f}$
6 Continue from 1 and let $f = f + 1$ until proper description of $\mathbf{y}_0$

Based on the extracted PCs and the scores, we can then obtain an estimation of a $t \times m$ regression matrix, $\hat{\mathbf{B}}$, for direct regression of a $1 \times m \times t$ feature slice of $\mathbf{X}(:,:,:)$, representing the features of a particular sequence over time on our quality vector $\mathbf{y}$. Hence, the quality estimation (2) can now be written as

$$\hat{\mathbf{y}} = \mathbf{1}\hat{b}_0 + \mathbf{X}\hat{\mathbf{B}} + \mathbf{e}. \tag{6}$$

The quality of unknown video sequences can be predicted similarly to (3), where the feature vector $\mathbf{x}_u$ is replaced by a corresponding feature slice $\mathbf{X}_u$. For a more detailed description of Tri-PLS1, we refer to [8].

3. HYBRID NO-REFERENCE METRIC

In this section we will extend our previously proposed H.264/AVC bitstream feature based no-reference metric proposed in [1] with pixel-based features, leading to a new hybrid no-reference metric. We will design two different metrics by analysing the data with Tri-PLS1 resulting in two corresponding models, one for each feature class. Then we combine the quality prediction results from the pixel-based features $\hat{y}_P$ with the prediction results from the H.264/AVC bitstream-based features $\hat{y}_B$ in order to gain an overall quality prediction $\hat{y}$. The concept of the metric is illustrated in Fig. 1 and will be discussed in the following subsections.

3.1. Bitstream Feature Extraction

In a first step we extract features from the H.264/AVC bitstream, describing the properties of the encoded video sequence. We assume in the following that the *byte stream* representing the Network Abstraction Layer (NAL) according to

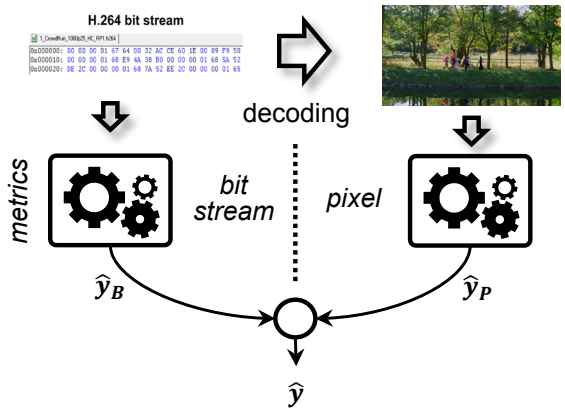


Fig. 1: Overview of the proposed hybrid metric: separate metrics for bitstream and pixel-based features and the combination of individual quality predictions $\hat{y}_B$ and $\hat{y}_P$ into an overall quality prediction $\hat{y}$

Annex B of the H.264/AVC standard is available and that any channel coding done for transmission has already been removed. We then parse those NAL units (NALU) containing information about the coded frames.

We extract the following features for each slice in the video sequence: slice type (*I, P, B*), kBits per slice (*BPS*), QP per slice (*QPA*), average and maximum motion vector length and motion vector error (*MV, MVMax, MVd, MVdMax*), percentage of intra, inter and skip coded macroblocks (*%Intra, %Inter, %Skip*), percentage of intra macroblocks with 16×16, 8×8 and 4×4 subdivision (*%I16x16, %I8x8, %I4x4*) and the percentage of inter macroblocks with 8×8 and 4×4 subdivision (*%P8x8, %P4x4*). While the large number of extracted features seem to imply an increased computational complexity, note that we only parsed the intrinsic parameters of the H.264/AVC bitstream. For more information about the bitstream features and their extraction, we refer to [1].

3.2. Pixel Feature Extraction

The pixel-based part of the proposed metric uses the following seven different no-reference features, that will be described shortly in this section: *blockiness*, *bluriness*, *activity*, *predictability*, *motion continuity*, *edge continuity* and *color continuity*. The first three features are *intra* features, describing aspects of the video quality with respect to one frame, and the later four are *inter* features, describing aspects of video quality with respect to changes between frames, based on the assumption that human observers prefer smooth transitions between neighbouring frames: predictability describing how well one frame can be predicted using only the previous frame, motion continuity measuring the smoothness of the motion, color continuity describing the color changes between two successive frames and edge continuity describing the change of edge regions between two successive frames.

Blurriness is measured as described in [9]. The algorithm measures the width of an edge and then calculates the blur by assuming that blur is reflected by wide edges, adjusted by a piecewise linear correction if the video contains high amount of fast motion. *Blockiness* is determined using the algorithm introduced in [10] by calculating the horizontal and vertical blockiness in the frequency domain. It compares the measured spectrum with the spectrum of a smoothed version of the frame. Spatial *activity* is derived from the amount of details that are described by the percentage of turning points along each line and row. It is part of the BTFR metric, included in [11]. For temporal *predictability*, an image is generated by motion compensation with a simple block matching algorithm. The current image and its prediction are then compared block by block. In order to avoid that single pixels dominate, both images are filtered using a Gaussian filter, followed by median filtering. The output of this process is the percentage of blocks that are not noticeably different. *Edge continuity* is determined by comparing the current frame with its motion compensated prediction using Edge-PSNR [12]. It reflects how much the structure of the image changes. *Motion continuity* assumes that most objects should follow a relatively smooth motion trajectory and that non-smooth motion trajectories may be caused by artefacts like jitter. Hence, two motion vector fields are calculated: between the current and the previous frame and between the current and the following frame. The difference is then leveraged to determine a measure for the motion continuity. Finally, *color continuity* is determined via the linear correlation between the color histograms of the current image and its prediction. It allows for gradient changes in color, but can indicate color artefacts e.g. color bleeding. For more information about these features, the algorithms to extract them and their application in no-reference metrics, we refer to [13].

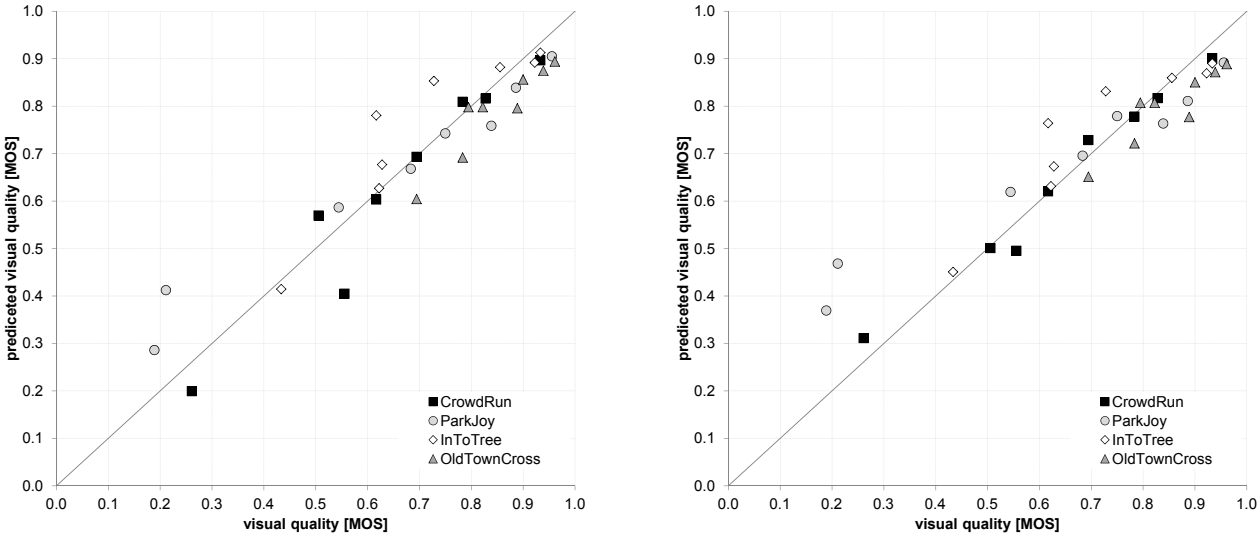
3.3. Combining Bitstream and Pixel Features

In the final step, we combine the quality predictions $\hat{y}_B$ for the bitstream-based model and $\hat{y}_P$ for the pixel-based model into one overall quality prediction $\hat{y}$:

$$\hat{y} = 0.85\hat{y}_B + 0.15\hat{y}_P. \tag{7}$$

The weighting parameters were determined during the model calibration process by linearly regressing the prediction results for the training data of both bitstream-based and pixel-based models on the visual quality and then averaging the regression coefficients over all models. Note, that these parameters are a fixed part of the overall metric and not the result of individual data fitting to each video sequence. Lastly, we apply a fixed sigmoid nonlinear correction to the prediction values $\hat{y}$ in order to emulate the nonlinear nature of the test results in subjective testing at the extrema of the scale. This correction is given as

$$\hat{y}_S = 1.0/(1 + e^{(-(\hat{y}-0.5)/0.2)}). \tag{8}$$



(a) Bitstream metric from [1]

(b) Hybrid metric

Fig. 2: Prediction results for bitstream-based metric and hybrid metric, both designed with Tri-PLS1

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

4. EVALUATION

In the evaluation of the proposed hybrid metric's prediction performance, we used the video sequences from the TUM 1080p25 data set for calibration and validation. This set consists of four sequences from the established SVT multi format test set in 1080p25 HDTV format, namely *CrowdRun*, *InToTree*, *ParkJoy* and *OldTownCross*. These sequences were encoded in H.264/AVC at four bitrate points ranging from 5.4 Mbit/s to 30 Mbit/s to represent a sufficient distribution of visual quality. The data set is available at [14] and for further information we refer to [15].

We evaluated our metric with a leave-one-out cross-validation. Therefore, we build four subsets of the available data set, each containing all but one of the video sequences previously introduced. We then train our metric on the data of the remaining three video sequences and use the resulting four models to predict the quality of the video sequence that wasn't included in the training phase. By separating the data into different sets for training and validation, we avoid getting overly optimistic prediction results.

5. RESULTS

The prediction results of the hybrid metric are presented in Fig. 2 and Table 1. Besides the Pearson and Spearman rank order correlation coefficents, we also provide the root mean squared error (RMSE) between predicted and actual visual quality. For comparison, we included the results from our bitstream-based no-reference metrics presented in [1] and [16], where we used both PLS1 and Tri-PLS1, respectively. Additionally, we include the results of two well-known full-reference video quality metrics: SSIM [17] and the VQM according to Annex D of ITU-T J.144 [11]. While both metrics are general purpose metrics and therefore not as tuned to H.264/AVC artefacts as the proposed metric, they still provide a good baseline comparison to the state-of-the-art in video quality metrics.

The results show that the proposed hybrid metric outperforms both our purely bitstream-based metrics in [1] and [16] slightly with respect to both the Pearson correlation and the Spearman rank order correlation, showing the benefit of the inclusion of pixel-based features into the metric. Also we can notice in Fig. 2 that while the RMSE might be slightly worse, for the sequences containing a larger amount of temporal variation, *CrowdRun* and *ParkJoy*, the prediction is getting even closer to the desired linear relationship, whereas those sequences with less temporal variation, *InToTree* and *OldTownCross*, do not benefit similarly by the inclusion of the additional information contained in the pixel-based features. This indicates that these additional features help us to model the temporal quality variation better. However, note in Fig. 2 that due to the lack of low quality data points in the training set, the prediction quality is worse at the lower end of the quality scale. Even tough the results are not directly comparable due to the use of different video sequences and data sets, we note that the proposed hybrid metric in this contribution also outperforms the hybrid metrics presented in [2–4] with respect to the Pearson correlation.

Table 1: Performance of the quality prediction

Metric	Pearson	Spearman	RMSE [(a)]
Bitstream metric			
PLS1 based [16]	0.93	0.95	0.08
Tri-PLS1 based [1]	0.94	0.93	0.07
Hybrid metric			
Tri-PLS1 based	0.95	0.94	0.08
PSNR	0.72	0.69	0.15
SSIM [17]	0.85	0.82	0.12
VQM Annex D of [11]	0.84	0.78	0.11

[(a)] After first order fitting for all comparison metrics, no fitting for PLS based metrics

6. CONCLUSION

We extended the design of video quality metrics with trilinear partial least square regression from a purely bitstream feature based metric into a hybrid metric by combining our previous metric with pixel-based features.

Our results show that the inclusion of the additional features gained from the decoded video sequences increase the prediction accuracy even further. In future work, larger data sets including different prediction structures and encoding settings should be considered to cover a larger quality range.

7. REFERENCES

[1] 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), Third International Workshop on*, Sep. 2011, pp. 49–54.

[2] K. Yamagishi, T. Kawano, and T. Hayashi, "Hybrid video-quality-estimation model for IPTV services," in *Global Telecommunications Conference (GLOBECOM)*, Dec. 2009, pp. 1–5.

[3] O. Sugimoto, S. Naito, S. Sakazawa, and A. Koike, "Objective perceptual video quality measurement method based on hybrid no reference framework," in *Image Processing (ICIP), 16th IEEE International Conference on*, Nov. 2009, pp. 2237–2240.

[4] M. Farias, M. Carvalho, H. Kussaba, and B. Noronha, "A hybrid metric for digital video quality assessment," in *Broadband Multimedia Systems and Broadcasting (BMSB), IEEE International Symposium on*, Jun. 2011, pp. 1–6.

[5] N. Staelens, I. Sedano, M. Barkowsky, L. Janowski, K. Brunnstrom, and P. Le Callet, "Standardized toolchain and model development for video quality assessment - The mission of the joint effort group in VQEG," in *Quality of Multimedia Experience (QoMEX), Third International Workshop on*, Sep. 2011, pp. 61–66.

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

[7] H. Martens and M. Martens, *Multivariate Analysis of Quality*. Wiley & Sons, 2001.

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

[9] 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.

[10] 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.

[11] *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.

[12] C. Lee, S. Cho, J. Choe, T. Jeong, W. Ahn, and E. Lee, "Objective video quality assessment," *SPIE Optical Engineering*, vol. 45, p. 7004, Jan. 2006.

[13] T. Oelbaum, C. Keimel, and K. Diepold, "Rule-based no-reference video quality evaluation using additionally coded videos," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 2, pp. 294–303, Apr. 2009.

[14] TUM Institute for Data Processing. (2012, Mar.) TUM 1080p25 Data Set. [Online]. Available: http://www.ldv.ei.tum.de/videolab/

[15] C. Keimel, J. Habigt, T. Habigt, M. Rothbucher, and K. Diepold, "Visual quality of current coding technologies at high definition IPTV bitrates," in *Multimedia Signal Processing (MMSP), IEEE International Workshop on*, 2010, pp. 390–393.

[16] 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), 18th IEEE International Conference on*, Sep. 2011, pp. 3325–3328.

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