SALIENCY BASED VIDEO QUALITY PREDICTION USING MULTI-WAY DATA ANALYSIS

Arne Redl, Christian Keimel, Klaus Diepold

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

ABSTRACT

Saliency information allows us to determine which parts of an image or video frame attracts the focus of the observer and thus where distortions will be more obvious. Using this knowledge and saliency thresholds, we therefore combine the saliency information generated by a computational model and the features extracted from the H.264/AVC bitstream, and use the resulting saliency-weighted features in the design of a video quality metric with multi-way data analysis. We used two different multi-way methods, the two dimensional principal component regression (2D-PCR) and multi-way partial least squares regression (PLSR) in the design of a no-reference video quality metric, where the different saliency levels are considered as an additional direction. Our results show that the consideration of the the saliency information leads to more stable models with less parameters in the model and thus the prediction performance increases compared to metrics without saliency information for the same number of parameters.

Index Terms— H.264/AVC, video quality metrics, 2D-PCR, PLSR, saliency

1. INTRODUCTION

Objective video quality metrics are still in the focus on the research of video quality, as subjective video quality evaluation is often expensive and time-consuming. Obviously, a complete model of the human visual system (HVS) would be the best basis for such a video quality metric. Unfortunately the HVS is a rather complex system that has not been understood sufficiently enough to build such a model. Another design concept is a data driven approach to predict subjective video quality as presented in our previous contributions [1] and [2], where we utilized two-dimensional principal component regression (2D-PCR) and multi-way partial least square regression (PLSR). These metrics are no-reference metrics that do not need the undistorted video and moreover include the temporal dimension of video without pooling into the video quality estimation process.

Although there is no general model of HVS, some aspects can be described very well, in particular the prediction of the human visual attention or saliency. This allows us to determine which parts of an image or video frame attracts the focus of the observer and thus where distortion will be more obvious. Hence, we can decide if features are more or less relevant for the perceived quality, providing us an additional information source in the visual quality assessment. In this contribution, we consider the saliency information and its different levels as an additional direction in the data analysis, providing our model with information about the importance of extracted features. As we aim at the design of applicable video quality metric, we use a computational saliency model instead of eye-tracking data, as eye-tracking data is usually only available for data from subjective testing, whereas the computational saliency model can be applied to any video sequence, especially in an automated video quality assessment environment, where subjective testing is not feasible.

In related work, Alers et al. have shown that there are significant differences in the saliency maps of still images, if the observers had to judge the quality compared to free viewing [3]. For videos, a similar result was reported by Alers et al. in [4], also confirmed by Le Meur et al. in [5]. But to exclude any of this influence, we chose a computational model for generating the saliency data. Some full reference metrics using saliency maps have been suggested so far, e.g. Engelke et al. [6], Feng et al. [7] and You et al. [7], supplemented by proposed no-reference metrics, e.g. Boujut et al. [8] and Zhu et al. [9]. All these metrics, however, generally apply some form of pooling for the saliency data either spatially or temporally in order to achieve a quality value for the whole video sequence. But pooling, either temporally or spatially may destroy some interdependencies and should thus be avoided.

This contribution is organized as follows: First we introduce the data sets and the feature extraction, before discussing the saliency information and how this information and the features are combined. After introducing the used data analysis methods, we present the results and conclude with a short summary.

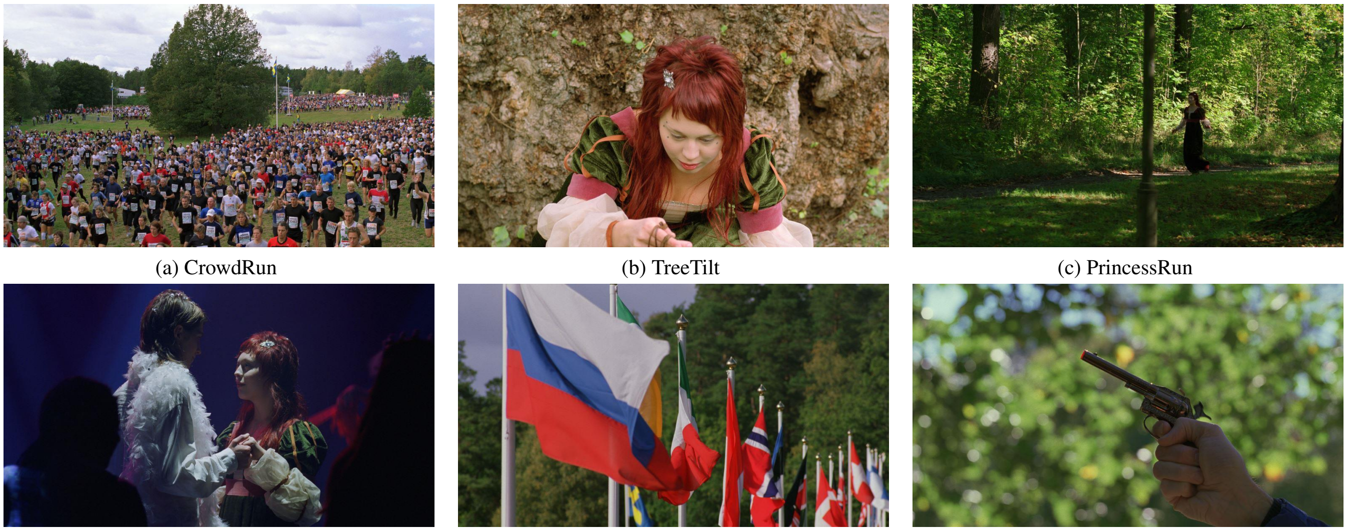
2. VIDEOS, FEATURES AND SALIENCY

In this section, we briefly introduce the used data set and how both the features and the salience information is extracted from the video sequences in the used data set.

2.1. Data Set

For the design and evaluation of the proposed video quality metrics we used a subset of the TUM High Definition Video Datasets. This dataset was generated in the ITU-R BT.500 [10] compliant video quality evaluation laboratory at the Institute for Dataprocessing at the Technische Universität München. We used the 1080p50 subset of the data set, consisting of five scenes from the well-known SVT multi format test set encoded with the reference implementation of H.264/AVC encoder, JM version 17.1, at 50 $\mathrm{frames/\,sec}$. Each scene has a length of 10 sec. This corresponds to 500 frames for the scenes *CrowdRun*, *TreeTilt*, *PrincessRun*, *DanceKiss* and 491 frames for the scene *FlagShoot* as shown in Fig. 1. All scenes were encoded at four quality levels to cover a large range of perceived quality, from bad to good visual quality. All in all, we have 20 different sequences with corresponding subjective visual quality as mean opinion scores (MOS) based on a discrete voting scale from 0 to 10. For more information we refer to [11] and the results of this dataset are also discussed in detail in [12].

Digital Object Identifier: 10.1109/QoMEX.2013.6603235



(a) CrowdRun

(b) TreeTilt

(c) PrincessRun

(d) DanceKiss

(e) Flags

(f) Shoot

Fig. 1: Sequences from the TUM HD Testsets (1080p50)

2.2. Feature Extraction

In order to weight the features extracted from the H.264/AVC bitstream with the saliency information, we obviously need information about the spatial location of the features in each video frame. Hence, we can not take the same approach as in [2], where we extracted the features only for each frame, but not on a sub-frame scale. Nevertheless, we will use similar features as in [2], but in contrast to [2] we do not pool them on a frame level. Using the modified H.264/AVC decoder from the Video Coding Expert Group (VQEG) [13], we are able to extract features from the H.264/AVC bitstream on a macro block level with exact knowledge of their spatial position in the frame. Thus we obtain one or more values for every macro block in the sequence for the following features:

- *Quantization parameter (QP)* with possible values $0 - 51$
- *Motion vector* one value for the absolute and one value for the difference vector, for each x and y direction
- *Skip flag* for blocks which are marked as skipped
- *Block type* possible types are I for intra and B or P for inter frame blocks
- *Block size* possible values are 16×16, 16×8, 8×16, 8×8 and additionally 4×4 for intra frames.

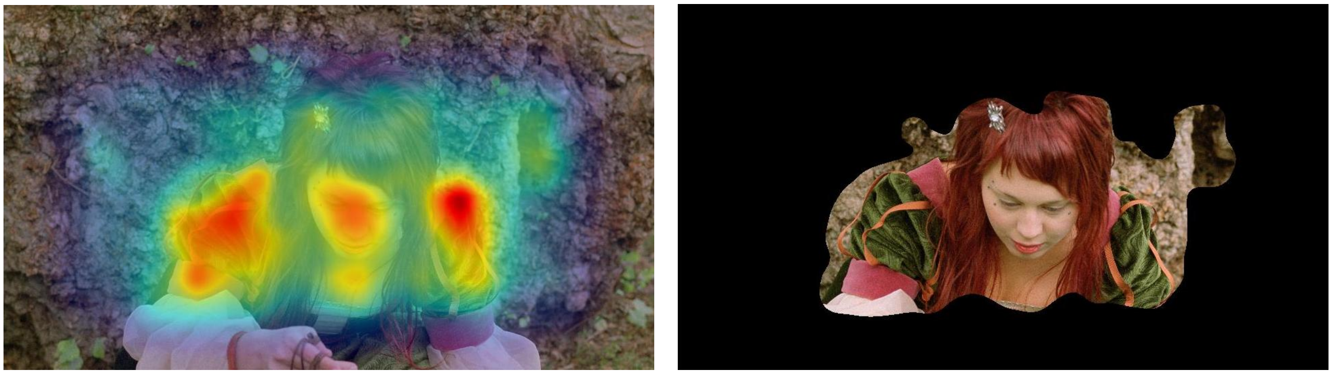
In the feature extraction we do not consider submacro blocks separately and therefore pool the motion vectors of each submacro block into the average motion vector of all motion vectors in a macroblock. Thus we obtain a data tensor $\underline{\boldsymbol{X}}' \in \mathbb{R}^{y \times x \times t}$ for every extracted feature, where x and y denote the spatial resolution in pixel and t denotes the temporal resolution in frames. For the QP and skip flag this feature tensor can be constructed straightforwardly: we obtain one tensor with all the QP values and a tensor with entries 0 or 1 for the skip flag, respectively. For the motion vectors we obtain four tensors, one tensor for each of vectors' direction at the given coordinates. For the block type and block size we obtain one tensor with the entries 0 or 1 for each block type and block size, respectively.

2.3. Saliency

The saliency information is based on a computational model proposed by Harel et al. [14]. This model was selected, because it is well known and understood, while performing very well compared with the data gained in eye-tracking experiments. The model provides a biological plausible bottom-up visual saliency model, working in roughly three stages: firstly, it generates feature maps concerning the intensity contrast, the mutual influence on the perception of the red/green and blue/yellow colour stimuli, and the local orientation information. Then in steps two and three the features are activated, normalized and combined, resulting in a saliency map. This saliency map has one entry for every pixel in every frame. Thus we can write this saliency map as a data tensor $\underline{\boldsymbol{S}} \in \mathbb{R}^{y \times x \times t}$ where y and x represent the spatial resolution and t the temporal resolution of a sequence, where each entry of $\underline{\boldsymbol{S}}$ is denoted as $s_{ijk} \in [0; 1]$. One advantage of using a computational model instead of data from an eye-tracking system is on the one hand the reproducibility for any desired videos, including videos not contained in a data set with eye-tracking results, on the other hand the proposed no-reference metric is thus also applicable for unknown videos, even if no eye tracking data is available. For more details on this saliency model we refer to [14, 15, 16]. An example of the resulting saliency map is given for a frame from the scene TreeTilt in Fig. 2: in Fig. 2a the frame is overlapped by a heat map and in Fig. 2b only the parts with a saliency value above 0.4 are shown.

3. COMBINING THE FEATURES WITH SALIENCY

After determining the saliency, we have one saliency weighting value for every pixel in the whole sequence, represented by $\underline{\boldsymbol{S}}$. Additionally, we obtain for each of the fourteen H.264/AVC bitstream features a data tensor $\underline{\boldsymbol{X}}' \in \mathbb{R}^{y \times x \times t}$ describing the spatial and temporal location of the corresponding feature values, where x, y and t represent the spatial and temporal resolution of the sequence. Assuming ten saliency thresholds $\hat{S} \in \{0; 0.1; ...; 0.9\}$, we can obtain for every entry in the data tensor $\underline{\boldsymbol{X}}'$ a new data tensor $\tilde{\underline{\boldsymbol{X}}}$



(a) Heat map

(b) Saliency threshold with $\hat{S} = 0.4$

Fig. 2: Visualizing the saliency of one frame in sequence *TreeTilt*

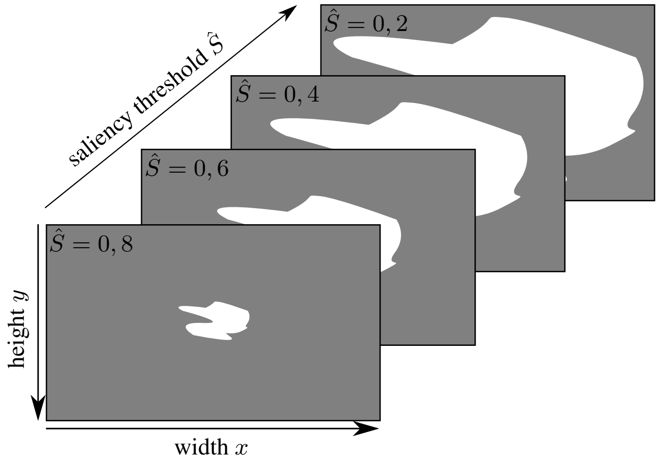


Fig. 3: Saliency thresholds

with the entries

$$\tilde{x}_{ijk} = \begin{cases} x'_{ijk} & \text{if } s_{ijk} \geq \hat{S} \\ 0 & \text{elsewhere} \end{cases} \tag{1}$$

for all $i \in [1; y]$, $j \in [1; x]$ and $k \in [1; t]$.

With the ten salience thresholds and the fourteen three-way data tensors $\underline{\tilde{\boldsymbol{X}}}$, this leads to fourteen four-way data tensors $\underline{\hat{\boldsymbol{X}}} \in \mathbb{R}^{y \times x \times t \times s}$, one for each feature. In order to avoid an additional direction in the data analysis, we replace the temporal dimension in 2D-PCR and PLSR by a direction describing the different saliency thresholds. Although data analysis methods for handling four-way data arrays exist, these methods have so far not received as much attention in literature as the special case of three-way arrays, already used and well understood in the context of video quality metrics. Hence, we decided to reduce the four-way array to a three-way array, by averaging all values from $\underline{\tilde{\boldsymbol{X}}}$ for every feature over the temporal direction. Note, that the saliency itself has not been pooled, but only the features. For each of the s saliency thresholds, the m features are then averaged over all pixels in the saliency threshold, resulting in fourteen average feature values per saliency threshold, corresponding to the number of extracted features. This is repeated for each of the n video sequences. Thus we get one value per sequence, feature and threshold, resulting in a new feature tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{n \times m \times s}$ for n sequences of the test set with m temporally pooled features and s saliency thresholds.

4. MULTI-WAY DATA ANALYSIS

In this section we briefly introduce the data analysis methods used to design the video quality metrics. For more information, we refer to [17, 18].

4.1. 2D-PCR

In [1] we presented an approach to the design of video quality metrics with two dimensional principal component regression (2D-PCR). 2D-PCR can be understood as an extension of the conventional principal component regression (PCR) to multi-way data[1]. Based on a feature tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{n \times m \times s}$ for the n video sequences in the training set we compute the average covariance matrix with

$$\boldsymbol{X}_{Cov} = \frac{1}{s} \sum_{k=1}^{s} \underline{\boldsymbol{X}}_{::k}^{\top} \underline{\boldsymbol{X}}_{::k}, \tag{2}$$

representing a measurement of the average temporal variation within $\underline{\boldsymbol{X}}$. Applying a PCA on $\boldsymbol{X}_{Cov}$ by performing a singular value decomposition (SVD) of $\boldsymbol{X}_{Cov}$ as

$$\boldsymbol{X}_{Cov} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{P}^{\top}, \tag{3}$$

we can then determine the scores array $\underline{\boldsymbol{T}} \in \mathbb{R}^{n \times g \times s}$ of $\underline{\boldsymbol{X}}$ with

$$\underline{\boldsymbol{T}}_{::k} = \underline{\boldsymbol{X}}_{::k} \boldsymbol{P} \quad \forall k \ \in [1; s]\,. \tag{4}$$

The scores $\underline{\boldsymbol{T}}_{::k}$ of each slice $\underline{\boldsymbol{X}}_{::k}$, representing the features for a given saliency threshold, are a projection of the slices $\underline{\boldsymbol{X}}_{::k}$ onto a subspace defined by the loadings $\boldsymbol{P}$, that explain the average covariance of all saliency thresholds best. Thus we expressed the original, saliency weighted features in $\underline{\boldsymbol{X}}_{::k}$ in terms of a new coordinate system given by $\boldsymbol{P}$. Note, however, that this coordinate system is only depending on the average covariance of the saliency thresholds, not on the variance within each saliency threshold. With the g largest

[1] For an in-depth discussion of tensors and their notation we refer to [18].

Eigenvalues of $\underline{\boldsymbol{T}}$ we extract a tensor $\underline{\boldsymbol{T}}_g$ with the first g lateral slices of $\underline{\boldsymbol{T}}$. We then get the prediction weights

$$\underline{\hat{\boldsymbol{C}}}_{::k} = \left(\underline{\boldsymbol{T}}_{g::k}{}^{\top}\underline{\boldsymbol{T}}_{g::k}\right)^{+}\underline{\boldsymbol{T}}_{g::k}{}^{\top}\boldsymbol{y} \quad \forall k \ \in [1;s] \tag{5}$$

where $\boldsymbol{T}^{+}$ denotes the More-Penrose pseudo-inverse and $\boldsymbol{y}$ the visual quality of the training set. In the original feature space we have now the weights in $\underline{\hat{\boldsymbol{B}}} \in \mathbb{R}^{m\times 1\times s}$ with

$$\underline{\hat{\boldsymbol{B}}}_{::k} = \boldsymbol{P}_g\underline{\hat{\boldsymbol{C}}}_{::k} \quad \forall k \ \in [1;s]\,. \tag{6}$$

Thus we gain the regression weights $\underline{\hat{\boldsymbol{B}}}$ for each saliency threshold with respect to the overall variation in all saliency thresholds. For a unknown sequence with the feature tensor $\underline{\boldsymbol{X}}_U \in \mathbb{R}^{1\times m\times s}$ we are now able to predict the subjective video quality by

$$\hat{\boldsymbol{y}}_k = \underline{\boldsymbol{X}}_{U_{::k}}\underline{\hat{\boldsymbol{B}}}_{::k} \quad \forall k \ \in [1;s]\,. \tag{7}$$

The vector $\hat{\boldsymbol{y}} \in \mathbb{R}^{\boldsymbol{1}\times \boldsymbol{s}}$ contains one value per saliency threshold and is then averaged to a scalar prediction value $\hat{y}$.

4.2. PLSR

An alternative regression method is the multi-way partial least squares regression (PLSR) an mulit-way extension of the well-known PLSR. It decomposes the feature tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{n\times m\times s}$ into scores $\boldsymbol{t}$ representing the n video sequences in the training set and loading weights $\boldsymbol{w}^m$ and $\boldsymbol{w}^s$, corresponding to the features and saliency, respectively.

In comparison to the 2D-PCR, we decomposed the three-way tensor $\underline{\boldsymbol{X}}$ directly into three components, one for each direction, whereas in the 2D-PCR although the regression was performed on the three-way array, the components were only extracted for a two-way array represented by the average covariance matrix $\boldsymbol{X}_{Cov}$. Thus we preserve more of the information in the three-way array with the multi-way PLSR. Another advantage compared to the 2D-PCR is that not only the variance of $\underline{\boldsymbol{X}}$ is explained but also the covariance of $\underline{\boldsymbol{X}}$ with $\boldsymbol{y}$.

For the three way data array $\underline{\boldsymbol{X}}$ we use the iterative trilinear PLS1 algorithm shown in Algorithm 1. This algorithm decomposes $\underline{\boldsymbol{X}}$ in its components $\boldsymbol{w}^m$ and $\boldsymbol{w}^s$, with $\boldsymbol{Z}$ as the matrix with the entries

$$z_{ms} = \sum_{i=1}^{n} y_n x_{nms}. \tag{8}$$

Algorithm 1: Trilinear PLS1 [19]

```
center $\underline{\boldsymbol{X}}$ and $\boldsymbol{y}$;
$\boldsymbol{y}_0 = \boldsymbol{y}$;
$\underline{\boldsymbol{X}}_0 = \underline{\boldsymbol{X}}$;
$g = 0$;
repeat
    Calculate $\boldsymbol{Z}$;
    Determine $\boldsymbol{w}_g^m$ and $\boldsymbol{w}_g^t$ by SVD of $\boldsymbol{Z}$;
    Calculate $\boldsymbol{t}_g$;
    $\boldsymbol{T} = \begin{pmatrix}\boldsymbol{t}_1 & \cdots & \boldsymbol{t}_g\end{pmatrix}$;
    $\boldsymbol{b}_g = \left(\boldsymbol{T}^{\top}\boldsymbol{T}\right)^{-1}\boldsymbol{T}\boldsymbol{y}_g$;
    $\boldsymbol{X}_{g+1} = \boldsymbol{X}_g - \boldsymbol{t}_g\boldsymbol{w}_g^m\left(\boldsymbol{w}_g^s\right)^{\top}$ and $\boldsymbol{y}_{g+1} = \boldsymbol{y}_g - \boldsymbol{T}\boldsymbol{b}_g$;
    Let $g = g + 1$;
until Proper description of $\boldsymbol{y}_g$;
```

Model	Saliency used	g	r_P	r_S	$RMSE$
2D-PCR	no	2	0.56	0.27	1.94
	yes	2	0.89	0.76	1.09
PLSR	no	1	0.45	0.61	2.13
	yes	1	0.90	0.78	1.04
SSIM	no	–	0.85	0.91	5.41

Table 1: Prediction performance for different number of used components g and saliency: Pearson correlation coefficient r_P, Spearman correlation coefficient r_S and root squared mean error $RMSE$

The score t_i for each samples is computed with the PCs with

$$t_n = \sum_{j=1}^{m}\sum_{k=1}^{s} x_{nms} w_i^m w_k^s. \tag{9}$$

Overall we obtain the weighting matrix $\hat{\boldsymbol{B}}$, which directly predicts the quality vector from a feature array $\underline{\boldsymbol{X}}_U$ of a unknown sequence similar to (7).

4.3. Preprocessing and Postprocessing

Both $\underline{\boldsymbol{X}}$ and $\boldsymbol{y}$ are preprocessed by centering in order to remove the average of every feature. Firstly we center the training data $\underline{\boldsymbol{X}}_T$ with

$$\underline{\boldsymbol{X}}_{T_{cent::s}} = \underline{\boldsymbol{X}}_{T::s} - \boldsymbol{1}\bar{\boldsymbol{x}}_{T_s} \tag{10}$$

for all s saliency thresholds. Unknown video sequences $\underline{\boldsymbol{X}}_U$ are centered by subtracting the feature average from the training set $\boldsymbol{1}\bar{\boldsymbol{x}}_{T_k}$. Similarly, we center the vector $\boldsymbol{y}$ containing the MOS with

$$\boldsymbol{y}_{cent} = \boldsymbol{y} - \boldsymbol{1}\bar{y}. \tag{11}$$

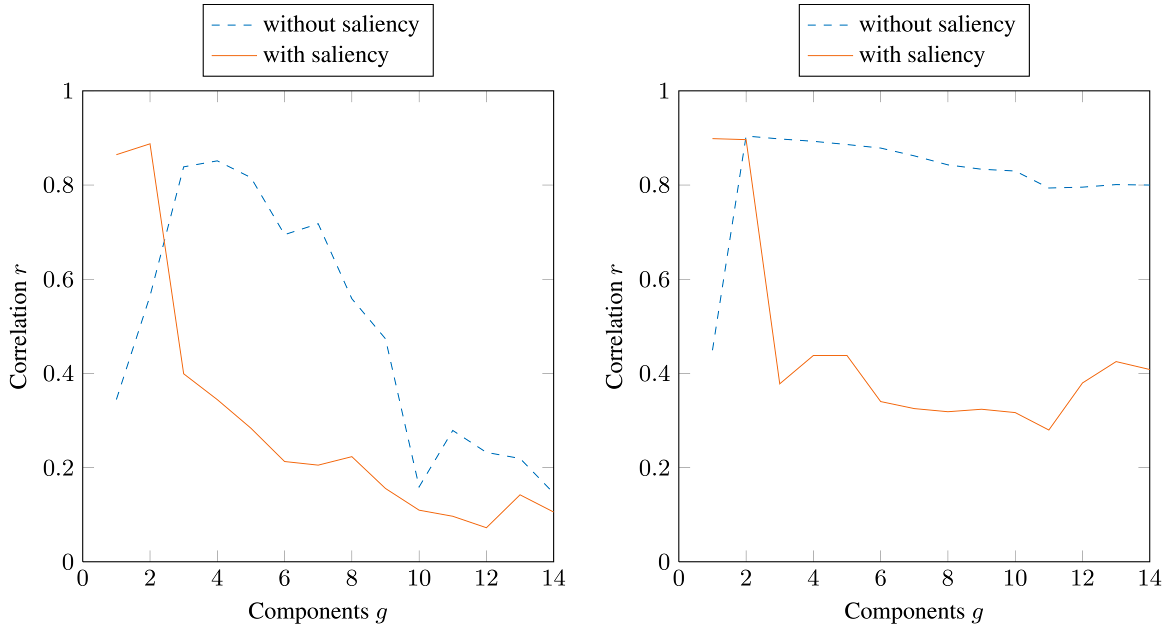
In order to avoid quality scores outside the expected quality range, we perform a correction step to limit them to the expected range. Additionally, we use this nonlinear function also emulate the nonlinear voting behaviour of test participants in the upper and lower regions of the voting scale. We use a sigmoid function

$$\hat{y} = \frac{a}{1 + \mathrm{e}^{-\frac{\hat{y}_M - b}{c}}} \tag{12}$$

where $\hat{y}_M$ are the values from the prediction model and $a = 1.0$, $b = 0.5$ und $c = 0.2$. Note, that this is a fixed part of the metric and the parameters in (12) are *not* depending on the actual data

4.4. Cross Validation

In order to avoid misleading results due to over-fitting it is necessary to use different video sequences in training and validation. But as the number of different sequences in the available datasets is limited and has to be used as efficiently as possible, we performed a leave-one-out cross validation. We always excluded all sequences with one specific content and obtained in this way five different training sets. The video sequences excluded in the training were then consequently used to validate the model built without the excluded video sequences.



(a) 2D-PCR

(b) PLSR

Fig. 4: Correlation coefficients r depending on the amount of components g

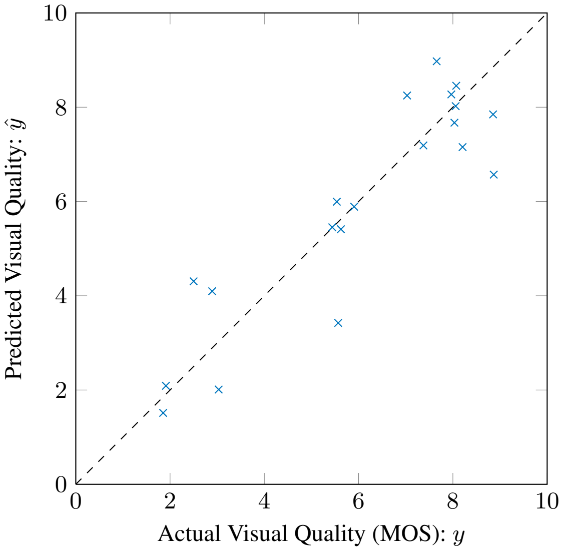


Fig. 5: Predicted vs. actually MOS for PLSR and $g = 1$

5. RESULTS

The result in Fig. 4 shows that we achieve a good Pearson correlation for the data with and without saliency. Similar prediction performance was achieved by Keimel et al. in [1] with other data sets, but similar features. Keimel et al. used features on a slice level and features regarding the whole sequence. Hence, this confirms that features extracted on a macro block level are able to provide similar information as features on a frame level with respect to the visual quality prediction. The second and probably more significant result is that using saliency, we achieve the same correlation as without saliency, but with significantly less components for 2D-PCR and PLSR. This provides mainly two advantages: its need less computational effort as we have to extract fewer components and it suggests a higher stability as less components are needed, by providing a more parsimonious model. The assumption is that the more components are needed, the more likely it is that most of the components only describe the noise in the data and not any latent structures. Hence the prediction performance for completely unknown video sequences is likely to suffer more for models that need a higher number of components for acceptable prediction results. It can also be shown in general that (linear) models with fewer components and thus a more parsimonious model on average have a smaller prediction error asymptotically [20]. The relationship between the number of components and the prediction performance is shown in Fig. 4 and Table 1. Noticeably, the trilinear PLSR provides a more comprehensive consideration of the saliency dimension compared to 2D-PCR and needs less components than 2D-PCR, confirming its theoretical advantage due to its better consideration of the three-way structure of the data. Using the saliency information, we achieve the best prediction performance with only one and two components for PLSR and 2D-PCR, respectively. For PLSR with $g = 1$ the resulting quality prediction is also shown in Fig. 5.

Also it compares very well to the de-facto standard in visual quality assessment, the full-reference SSIM [21] as shown in Table 1. Only with respect to the Spearman rank order correlation the prediction performance of our approach is lower than for SSIM. Because of the lack of freely available saliency-based no-reference quality prediction models, we were unfortunately not able to perform a comparison to similar prediction models. Furthermore we need only about 2 % of the original amount of the input data due to the temporal pooling of the features and thresholding of the saliency. Using the saliency information the input data array is reduced in our case from $\underline{\boldsymbol{X}} \in \mathbb{R}^{n\times m\times 500}$ to $\underline{\boldsymbol{X}} \in \mathbb{R}^{n\times m\times 10}$.

6. CONCLUSIONS

We combined saliency-weighed H.264/AVC bitstream features with multi-way data analysis methods to design video quality metrics. Instead of saliency gained in eye-tracking experiments, we used a computational model for the saliency, allowing for the application of the designed metric in a real-life environment for any video sequences.

The results show that the inclusion of the saliency information leads to video quality models with increased stability due the use of less components. Additionally, we have shown that the H.264/AVC bitstream features extracted on a macroblock level delivers similar results to the features extracted on a frame or slice level.

In future work, we plan to examine this new approach for more and bigger data sets. Furthermore we intend to investigate the use of four-way data analysis methods in order to avoid the temporal pooling.

7. REFERENCES

[1] Christian Keimel, Martin Rothbucher, Hao Shen, and Klaus Diepold, "Video is a Cube," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 41–49, 2011.

[2] Christian Keimel, Julian Habigt, Manuel Klimpke, and Klaus Diepold, "Design of no-reference video quality metrics with multiway partial least squares regression," in *Third International Workshop on Quality of Multimedia Experience (QoMEX), 2011*, Piscataway N.J., 2011, pp. 49–54, IEEE.

[3] Hani Alers, Lennart Bos, and Ingrid Heynderickx, "How the task of evaluating image quality influences viewing behavior," in *Third International Workshop on Quality of Multimedia Experience (QoMEX), 2011*, Piscataway N.J., 2011, pp. 167–172, IEEE.

[4] Hani Alers, Judith A. Redi, and Ingrid Heynderickx, "Examining the effect of task on viewing behavior in videos using saliency maps," in *Human vision and electronic imaging XVII*, Bernice Ellen Rogowitz, Thrasyvoulos N. Pappas, and Huib de Ridder, Eds., Bellingham, 2012, SPIE and IS&T.

[5] Olivier Le Meur, A. Ninassi, Patrick Le Callet, and Dominique Barba, "Overt visual attention for free-viewing and quality assessment tasks," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 547–558, 2010.

[6] Ulrich Engelke, Marcus Barkowsky, Patrick Le Callet, and Hans-Jurgen Zepernick, "Modelling saliency awareness for objective video quality assessment," in *Second International Workshop on Quality of Multimedia Experience (QoMEX), 2010*, Piscataway N.J., 2010, pp. 212–217, IEEE.

[7] Xin Feng, Tao Liu, Dan Yang, and Yao Wang, "Saliency based objective quality assessment of decoded video affected by packet losses," in *15th IEEE International Conference on Image Processing, 2008*, Piscataway N.J., 2008, pp. 2560–2563, IEEE.

[8] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet, "A metric for no-reference video quality assessment for HD TV delivery based on saliency maps," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, July, pp. 1–5.

[9] Lin Zhu, Li Su, Qingming Huang, and Honggang Qi, "Visual saliency and distortion weighting based video quality assessment," in *Advances in Multimedia Information Processing - PCM 2012*, Weisi Lin, Dong Xu, Anthony Ho, Jianxin Wu, Ying He, Jianfei Cai, Mohan Kankanhalli, and Ming-Ting Sun, Eds., vol. 7674 of *Lecture Notes in Computer Science*, pp. 546–555. Springer Berlin Heidelberg, 2012.

[10] ITU-R (Radiocommunication Sector of ITU), "Recommendation ITU-R BT.500-12 (09/2009): Methodology for the subjective assessment of the quality of television pictures," 2009.

[11] Christian Keimel, Arne Redl, and Klaus Diepold, "The TUM High Definition Video Datasets," in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX), 2012*, Piscataway N.J., 2012, pp. 97–102, IEEE.

[12] Arne Redl, Christian Keimel, and Klaus Diepold, "Influence of viewing device and soundtrack in HDTV on subjective video quality," in *Image quality and system performance IX*, Frans Gaykema and Peter D. Burns, Eds., Bellingham, 2012, SPIE and IS&T.

[13] Video Quality Experts Group (VQEG), "Modified JM H.264/AVC Codec," .

[14] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*. 2007, pp. 545–552, MIT Press.

[15] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[16] Laurent Itti and Christof Koch, "Computational modeling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[17] Harald Martens and Tormod Næs, *Multivariate Calibration*, John Wiley & Sons, Chichester and, 1989.

[18] Age K. Smilde, Rasmus Bro, and Paul Geladi, *Multi-way analysis with applications in the chemical sciences*, John Wiley & Sons, Chichester, 2005.

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

[20] Mary Beth Seasholtz and Bruce Kowalski, "The parsimony principle applied to multivariate calibration," *Analytica Chimica Acta*, vol. 277, no. 2, pp. 165 – 177, May 1993.

[21] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.