J. Vis. Commun. Image R. 43 (2017) 173–184

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

No-reference pixel based video quality assessment for HEVC decoded video ☆

Xin Huang, Jacob Søgaard, Søren Forchhammer *

DTU Photonics, Technical University of Denmark, Ørsteds Plads 343, 2800 Kgs Lyngby, Denmark

ARTICLE INFO

Article history:
Received 1 December 2015
Revised 15 November 2016
Accepted 1 January 2017
Available online 4 January 2017

Keywords:
HEVC analysis
No-reference
Video quality assessment
Machine learning
Elastic net

ABSTRACT

This paper proposes a No-Reference (NR) Video Quality Assessment (VQA) method for videos subject to the distortion given by the High Efficiency Video Coding (HEVC) scheme. The assessment is performed without access to the bitstream. The proposed analysis is based on the transform coefficients estimated from the decoded video pixels, which is used to estimate the level of quantization. The information from this analysis is exploited to assess the video quality. HEVC transform coefficients are modeled with a joint-Cauchy probability density function in the proposed method. To generate VQA features the quantization step used in the Intra coding is estimated. We map the obtained HEVC features using an Elastic Net to predict subjective video quality scores, Mean Opinion Scores (MOS). The performance is verified on a dataset consisting of HEVC coded 4 K UHD (resolution equal to 3840 × 2160) video sequences at different bitrates and spanning a wide range of content. The results show that the quality scores computed by the proposed method are highly correlated with the mean subjective assessments.

1. Introduction

Video Quality Assessment (VQA) has become an increasingly important issue in recent years, since the amount of video content compressed and transmitted over a wide range of devices from smart phones to large screen displays has increased significantly. As stated in [1], the latest video coding standard, High Efficiency Video Coding (HEVC) can provide approximately a 50% bit-rate savings for equivalent perceptual quality relative to the performance of prior standards, especially for a high-resolution video.

Human viewers are essentially the most reliable evaluator of video quality, rather than conventional measurements based on the mean-square error, for instance PSNR. Subjective scores or Means Opinion Scores (MOS) can be used to assess the video quality [2]. However, conducting subjective tests with human viewers requires a well-controlled test environment, which is resource demanding and infeasible to apply in many practical applications. Therefore, different ways of objective VQA are necessary depending on the target application and corresponding limitations.

Objective VQA can be divided into three main categories: Full-Reference (FR) [3], Reduced-Reference (RR) [4] and No-Reference (NR) [13] quality assessment. NR quality assessment is used in scenarios where the original video is unavailable. This type of VQA can be further divided into Pixel-Based (PB) methods and Bitstream-Based (BB) methods, depending on the processing level from where information is extracted. NR PB methods process the pixels of the decoded video with no access to the bitstream whereas NR BB methods extract the necessary information directly from the encoded bitstream. In many applications, it is not possible or convenient to access the video bitstream, e.g. in video forensics or if an encrypted broadcast stream is received by a set-top box that outputs the decoded video. Therefore, in this paper we focus on the NR PB VQA problem, where the reference signal is not available and with no access to the bitstream. In this setting we propose a novel VQA solution for HEVC video.

Objective NR Image Quality Assessment (IQA) and VQA have received much attention in the research community in recent years. Some state-of-the-art NR IQA methods are presented in [5–11]. However, NR IQA methods are often not applicable to video, since the training or modeling used in the methods are only valid for images. A review and classification of recent methods for NR VQA is presented in [12].

Some prior NR PB work for VQA is presented in [13–20]. In general, in [13–18] the approach is to measure the video quality through an analysis of the quantization step and/or a PSNR estimation for MPEG-2 and H.264/AVC coded video. In [14], an H.264/AVC based NR PSNR estimation is presented. In [15], a PB method for

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.

E-mail addresses: jsog@fotonik.dtu.dk (J. Søgaard), sofo@fotonik.dtu.dk (S. Forchhammer).

http://dx.doi.org/10.1016/j.jvcir.2017.01.002

estimation of the Quantization Parameter (QP) and motion vectors in H.264/AVC videos (without enabling in-loop De-Blocking filter) is described. In [16], a Discrete Cosine Transform (DCT) based NR video quality prediction model is described to analyze the statistics of H.264/AVC compressed videos. In [13,17,18], a series NR PB methods for Intra frames analysis in MPEG-2 video and H.264/AVC video (with De-Blocking enabled) are presented. In [19], the authors present a general NR PB VQA based on features extracted using a 3D shearlet transform, a convolutional neural network, and linear regression. It is tested on a number of distortion types and while having a good performance in general, it does not perform better than PSNR for HEVC content in the test reported. In [20], a NR VQA method for evaluating the impact of frame freezes is presented.

In this paper, a novel NR PB VQA method for the latest HEVC video coding is proposed and evaluated on UHD videos. Related HEVC VQA work, but for NR BB PSNR estimation is presented in [21]. In [22], a NR BB VQA method for H.264/AVC and HEVC estimating the FR IQA score SSIM [23] is presented. Different from [21,22], the proposed method evaluates video quality without bitstream access. Furthermore, the goal of the video quality assessment in this work is to estimate subjective video scores rather than objective measurements like PSNR in [21] or SSIM in [22]. In [24] we presented our initial work on VQA of HEVC videos by codec analysis, where the focus was on the BB approach.

As stated in [12], most existing NR methods are designed for cases where the quality difference is rather clearly visible. These methods might not work very well in scenarios where the source material is of very high quality and degradations between different quality levels therefore can be harder to perceive. This is often the case with HEVC videos in UHD resolutions and motivates the need for the proposed method. Besides compression HEVC videos might be degraded by e.g. packet loss in transmission systems. Packet loss for HEVC videos and the difference compared to H.264/AVC are investigated in [25], but are not considered in this work.

Without accessing the bitstream, the proposed method is able to estimate the position of the Intra frames and the corresponding QP, based on HEVC decoded videos. For inter frames we shall only perform spatio-temporal analysis on the decoded video. Extending the codec analysis of inter frames could also be possible, but it would be much more complex and is not considered in this work. Even though the codec analysis part of the proposed method does not consider inter frames, good results are obtained as reported in Section 6.

Compared to previous work related to H.264/AVC [13,18], this work focus on VQA according to the characteristics of HEVC. Also, contrary to the early work on HEVC VQA presented in [24], this paper focuses on the PB approach. The novelties of this work are introduced to achieve PB analysis of HEVC and they include: generation of the residuals after intra prediction based on the HEVC quad-tree structure, a coefficient model for the new transform and the filter introduced in HEVC, and a joint frame and Coding Tree Unit (CTU) level analysis scheme for QP estimation. Additionally, to the best of our knowledge, these are the first NR PB VQA results for estimating subjective video scores on HEVC sequences. We have tested on UHD sequences. The great increase in encoder options presents a serious challenge to the PB analysis. The research challenge has been to develop a scheme to analyze not all, but sufficient information about the increased number of modes, etc. in HEVC, thus striking a balance between performance and complexity.

The framework of the combined scheme consists of three main steps: First, HEVC intra prediction is performed on the decoded video to get an estimate of the transform coefficients of Intra frames. Secondly, using a joint-Cauchy distribution model of the transform coefficients after De-Blocking (DB) and Sample Adaptive Offset (SAO) filtering, a novel PB HEVC analysis method is applied to estimate the frame level quantization step and for Intra frame detection. Finally, the results of the HEVC analysis are mapped to a video quality score using machine learning (specifically the Elastic Net). An overview of the proposed method can be seen in Fig. 1. We assume that the QP is high enough to see the effects on the distribution of the transform coefficient and we therefore only consider HEVC videos with $\mathrm{QP} \in [20, \ldots, 51]$, which is a common range for general encoding. The lowest QP value in the evaluation dataset is equal to 19.

The paper is organized as follows. In Section 2, we introduce the basics about HEVC intra coding and rate control. This is followed by the details on how to generate an estimate of the transform coefficients based on processing of the decoded video. In Section 3, we elaborate on the proposed model for the distribution of the transform coefficients in HEVC videos. The estimation of frame level QP and the GOP size of HEVC is based on this model and is described in Section 4. In Section 5, we explain how we generate the features from the video codec analysis and how they are mapped to a video quality score by machine learning. Finally, we report the experimental results in Section 6.

2. HEVC basics

In this paper, all of the analysis is based on the luminance values of the decoded frames. In order to get an estimate of the transform coefficients of the intra frames in the HEVC encoder with no access to the bitstream, HEVC intra prediction shall be mimicked on the decoder side. Therefore, we start this section by reviewing the related basics of HEVC intra coding [1,26]. Thereafter, we briefly review relevant parts of the rate control scheme [27] in HEVC reference software (HM 11.0), which restricts the variations of QP values as is often the case for video rate control. Thus we restrict ourselves to the HEVC basics related to our approach and refer to [1] for an overview of HEVC.

2.1. HEVC intra coding

A fundamental difference between HEVC and previous standards is the usage of a Coding Tree Unit (CTU) based on a quad-tree structure. This is a flexible mechanism for subdividing a picture into different block sizes for prediction and residual coding. Conceptually speaking, there are three kinds of blocks defined in HEVC: Coding Unit (CU), Prediction Unit (PU), and Transform Unit (TU). CUs can be in intra, inter or skip mode with size variations from 8×8 to 64×64. For intra coding, the prediction mode is selected for each PU, while a CU can be divided into four square PUs, or treated as a single PU. Each PU can be split into square TUs following a quad-tree structure. All TUs within one PU share the same intra prediction mode. TU size is varying from 4×4 to 32×32. In practical implementations, the TU size often equals the PU size. In such cases, the TU becomes the basic unit for prediction and transform in HEVC intra coding. There are DC prediction, planar prediction and 33 directional intra predictions, in total 35 prediction modes in HEVC intra. For an $N \times N$ block, at most $4N + 1$ local neighboring pixels are used as reference samples. After prediction, transform coding of the prediction residual is applied. Different from prior standards, an alternative integer transform derived from the Discrete Sine Transform (DST) [28] is applied for intra coding with TU size 4×4. For other cases, the transform matrices are derived from the Discrete Cosine Transform (DCT) basis functions.

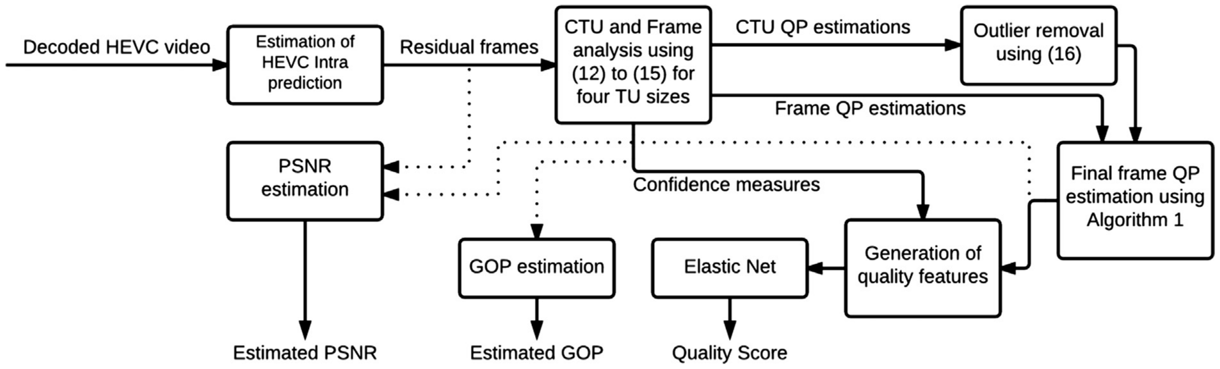

Decoded HEVC video
Estimation of HEVC Intra prediction
Residual frames
CTU and Frame analysis using (12) to (15) for four TU sizes
CTU QP estimations
Outlier removal using (16)
Frame QP estimations
Final frame QP estimation using Algorithm 1
PSNR estimation
Confidence measures
Generation of quality features
GOP estimation
Elastic Net
Estimated PSNR
Estimated GOP
Quality Score

Fig. 1. Flowchart of the proposed method. (Dotted lines indicate information used for GOP and PSNR estimation.)

In HEVC two in-loop filters are applied to the reconstructed pixels, namely the DB and SAO filter. The DB filter is similar to the DB filter in H.264/AVC, though more conditions are used in HEVC to decide where to apply the DB filter [1]. The SAO filter is applied adaptively to all reconstructed pixel values satisfying certain conditions and generally improves the visual quality of the signal in both smooth areas and around edges [29].

2.2. Rate control

Rate control plays an important role in video coding. Even though, the format and syntax related to rate control are specified, the specific rate control is not normative for most video coding standards. Rate control is applied to meet the bit allocation budget by setting appropriate QP values for each coding unit of the video. An R-λ (Rate versus Lagrange multiplier) model [27] based rate control scheme is adopted in the HEVC reference software (HM 11.0). The QP value can be set on a frame or CTU level. Our analysis, modeling, and QP estimation will focus on the CTU level, thus a typical CTU level rate control scheme presented in [27] is revisited in this section. This control scheme influences and constrains the QP value of each individual CTU to minimize quality variation. It can be exploited by the final QP estimation to remove outliers as shown in Fig. 1 and as detailed in Section 4. As stated in [27], the target bits T_{CurrCU} of each CTU is determined by:

$$T_{CurrCU} = \frac{T_{CurrPic} - Bit_{header} - Coded_{Pic}}{\sum_{NotCodedCTUs} \omega_i} \cdot \omega_{CurrCTU} \qquad (1)$$

where $T_{CurrPic}$ is the target number of bits for the current picture, Bit_{header} is the estimated bits of all headers, and ω is the weight of each CTU, estimated by the prediction error of the collocated picture in the previous coded picture. To achieve the target bits in terms of bits-per-pixel *bpp*, a corresponding λ has to be determined based on

$$\lambda = \alpha \cdot bpp^{\beta} \qquad (2)$$

where α and β are initialized to 3.2003 and -1.367, respectively [27]. After encoding one CTU, the real encoded *bpp* and real used λ value are used to adaptively update the values of α and β [27]. The QP value is determined according to the λ value by [27]

$$QP = 4.2005 \cdot \ln \lambda + 13.7122. \qquad (3)$$

To keep the quality consistent, the λ value and the determined QP value are clipped to a narrow range:

$$\lambda_{lastCTU} \cdot 2^{\frac{-1.0}{3.0}} \leqslant \lambda_{currCTU} \leqslant \lambda_{lastCTU} \cdot 2^{\frac{1.0}{3.0}} \qquad (4)$$

$$QP_{lastCTU} - 1 \leqslant QP_{currCTU} \leqslant QP_{lastCTU} + 1. \qquad (5)$$

3. Estimation and modeling of coefficients

In order to estimate and reproduce the transform coefficients of decoded intra frames as outlined in Section 2.1, we use a simplified HEVC CTU structure in the reproduction step. The CTU is fixed to 64 × 64 as the basic processing unit in the proposed scheme. At the HEVC encoder, there is only one best partition available for each CTU in the estimated residual frames, but which one is unknown in our case. Therefore, we apply four candidate CU sizes, 64 × 64, 32 × 32, 16 × 16, and 8 × 8 separately on each CTU of the decoded frame to be analyzed. The whole encoded frame is a mixture of CUs with different sizes. However, due to the huge number of combinations, we limit or reproduction to constant CU sizes.

In each of the four CU candidate frames introduced above, the CU is divided into four squared partitions to match the actual TU size from 32 × 32 to 4 × 4 as in the encoder. We consider each of the four different block sizes: 32 × 32, 16 × 16, 8 × 8, and 4 × 4. The best intra prediction mode is selected based on the decoded frames by using the Sum of Absolute Differences (SAD) as the criteria. This procedure is performed for each CTU of the decoded frames and results in four residual frames, one for each partition. As in the HEVC encoder, DST is only applied on 4 × 4 blocks, otherwise DCT is applied. After the transforms, all the coefficients of each residual frame are collected for further analysis and modeling at the CTU level as presented below in Sections 3.1 and 4.

3.1. Modeling of noisy coefficients

As stated in [30], the distribution of transform coefficients at the encoder before quantization can be accurately modeled by a single Cauchy distribution with zero mean. However, different from previous work as in [18,30] for H.264, our analysis is based on the luminance values of the decoded frames of HEVC. The quantization parameters, DB and SAO filters, the variations of transform block size (4 × 4 to 32 × 32) and the newly introduced DST in HEVC all highly affect the distribution of the obtained transform coefficients. The great increase in encoder options presents a serious challenge when there is no access to the bitstream.

We choose a mixture of Cauchy distributions as in [18]

$$f(x, qs) = \sum_{i=0}^{N} w_i p_i(x, qs, \gamma_i) \qquad (6)$$

where $N + 1$ is the number of distributions in the mixture model for signal x, qs denotes the quantization step, i is an index and integer denoting different multiplications of the quantization step, γ_i is the γ parameter in the ith Cauchy distribution, w_i is the corresponding weight, and p_i is the Cauchy distribution

$$p_i(x, qs, \gamma_i) = \gamma_i / (\pi \cdot ((x - i \cdot qs)^2 + \gamma_i^2)). \qquad (7)$$

The relationship between the quantization step and the QP value is given by:

$$qs(QP) = (2^{1/6})^{QP-4}. \tag{8}$$

We train our model using HEVC encoded versions of the reference video sequences from the LIVE dataset [31] at four different bitrates (for more detailed settings we refer to Section 6). It is well-known that the distributions of the transform coefficients at the encoder side decline rapidly with increasing distance from zero [30]. We observed in our training dataset that the distributions of the noisy decoded transform coefficients decline even more rapidly with increasing distance from zero, and the higher the quantization parameter, the wider the Cauchy distributions are. Motivated by these observations when using the true QP values of the training set, the γ_i parameter of the Cauchy distributions is defined as a first order polynomial to model the transform coefficients given by different partitions stated in Section 2 and to avoid overfitting:

$$\gamma_i(QP) = a_i + b_i * QP. \tag{9}$$

For each Cauchy distribution, $i \cdot qs$ equals the mean of the distribution. Thus, the Cauchy distribution p_i with $i = 0$ has zero mean, and it is therefore much harder to see the effect of different quantization steps as opposed to distributions with $i > 0$. Based on this observation, we discard the contribution from p_0. As expected, the weights decline rapidly with increasing distance to zero, so we only consider $i = \{1,2\}$. For $i = \{1,2\}$, the parameters in (9) are fitted separately for different partition size (from 4×4 to 32×32). The reason is that different partition sizes can introduce different distributions, since e.g. DST is applied for intra coding with TU size 4×4 while DCT is applied for the other cases. The parameters fitted for γ_i given in Table 1 are used. They are empirically based on the LIVE dataset [31], which was used as a training set. The weights in the mixture also depend on the quantization scale. Based on our empirical observations as seen in Fig. 2, w_1 was set to

$$w_1 = 0.006QP + 0.73 \; for \; QP < 45$$

$$w_1 = 1 \; for \; QP > 44 \tag{10}$$

while $w_2 = 1 - w_1$. It may be observed that w_2 is of much smaller value than w_1. However, setting $N = 2$ instead of $N = 1$ in (6) lead to better training accuracy in our implementation.

To determine the distribution of coefficients, a weighted sum of the coefficients is calculated in the analysis part of our approach (see Fig. 1), where a modified mixture of two Cauchy distributions is used as a weighting function. Since the peaks of the Cauchy distributions are lower for high QP parameters, an increased weight of the coefficients at higher quantization steps than coefficients at lower quantization steps, is defined. Therefore, a modified version of the distributions in the mixture is used:

$$\hat{P}_i(x, QP) = \begin{cases} C_p, & \begin{array}{l} if \; x = \lfloor i \cdot qs(QP) \rfloor \\ or \; x = \lceil i \cdot qs(QP) \rceil \end{array} \\ p_i(x, qs(QP), \gamma_i(QP)), & otherwise \end{cases} \tag{11}$$

where C_p (equal to 0.5 for $i = 1$ and equal to 0.7 for $i = 2$) is a constant value to balance the peak value with different quantization steps. It is obtained empirically by evaluating all the maximum p_i values over all quantization steps and different partition scenarios. The weighting function can be written as a weighted mixture of (11):

$$\hat{f}(x, QP) = \sum_{i=1}^{2} w_i(QP)\hat{P}_i(x, QP) \tag{12}$$

where the weights $w_i(QP)$ were empirically found to be in the range of $[0.85, 1]$ for w_1, while $w_2 = 1 - w_1$ as described above. Examples of the proposed weighting function for QP = 24 and 30 (where $qs(QP) = 10.08$ and 20.16) are shown in Fig. 3.

4. Intra frame analysis

Based on the proposed weighting function for the coefficients, the quantization steps for HEVC Intra coded frames are estimated as presented in this section. The QP estimation is not the final goal of the analysis, but calculating relevant features are used in the prediction of the video quality score as described in Section 5.1. If there is prior knowledge about the positions of the Intra frames, the QP analysis can be limited to Intra frames only. Otherwise the proposed analysis needs to be carried out on every frame to detect the position of Intra frames as described in Section 4.2. An overview of the method can be seen in Fig. 1.

4.1. QP estimation

Different from previous work on H.264/AVC [18], we need to apply the QP estimation for HEVC on four different residual frames (one for each partition as discussed in Section 2). Other novelties of the QP estimation for HEVC include a new response function for QP estimation, an analysis both at CTU level and frame level, a CTU level QP processing criteria and a newly designed scheme for final QP estimation. For each decoded HEVC frame, a response for QP $\in [20, \ldots, 51]$ is calculated based on the coefficients of each residual frame, at CTU and frame level, respectively. The response value reflects how well the actual distribution of the coefficients in a CTU or a frame matches the expectations for different QP values.

$$R(QP) = qs(QP) \sum_{j=0}^{M_x} H_j \hat{f}(x_j, QP) \tag{13}$$

where M_x denotes the index for the maximum of the quantized values x_j, H_j is the number of quantized values equal to x_j, and $\hat{f}$ is given by (12).

The multiplication with qs in (13) is done to counteract the overall exponentially decaying trend of $R(QP)$, which is due to the fact that the original coefficient values can be modeled as a Cauchy distribution with zero mean. However, even after including the $qs(QP)$ in (13), the function might still have an overall linear decreasing trend. Since we are interested in the strongest relative response, the response function $R(QP)$ is fitted with a polynomial $G(QP)$ of degree 4, and the estimation of QP is obtained with the strongest relative response as shown below. The chosen degree 4 of the fitting polynomial $G(QP)$ is a compromise between performance and complexity. Based on our test, it gives a too noisy relative response if a lower degree fitting polynomial as in [18] is employed for HEVC analysis, which in turn leads to an inaccurate QP estimation. An example of the response function $R(QP)$, the fitting polynomial $G(QP)$ and the corresponding QP with strongest relative response are shown in Fig. 4, using

$$G(QP) = g_4QP^4 + g_3QP^3 + g_2QP^2 + g_1QP + g_0 \tag{14}$$

$$q = argmax_{QP}(R(QP) - G(QP)) \tag{15}$$

where parameters g_i are found by a least squares fit of $G(QP)$ to $R(QP)$. Such relative response is applied at one individual CTU level and frame level, respectively. A matrix of QP predictions $[\widehat{QP}]^i_{M \times N}$ is obtained at the CTU level, where $\widehat{QP}^i_{(m,n)}$ denotes the estimation with coordinates (m,n), $0 \leqslant m \leqslant M, 0 \leqslant n \leqslant N$. $\widehat{QP}^i_{frame}$ denotes the frame level estimation. $i \in [1,4]$ is an index of the prediction size

Table 1
The fitting parameters of γ_i for $i = \{1,2\}$.

	4 × 4	8 × 8	16 × 16	32 × 32
$\mathbf{a_i}$	{8.69, 1.31}	{9.29, 0.89}	{6.25, −0.66}	{4.67, −1.85}
$\mathbf{b_i}$	{−0.09, 0.01}	{−0.06, 0.06}	{0.06, 0.17}	{0.09, 0.22}

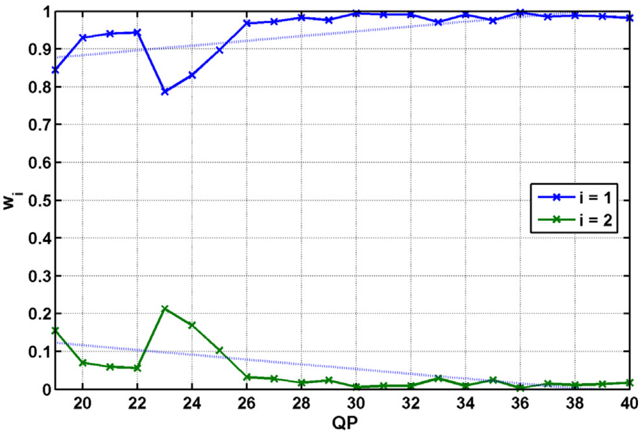


Fig. 2. Mean weights for the mixture of Cauchy distributions when fitted to the training video data for $i = \{1,2\}$.

from 4 × 4 to 32 × 32 in both cases. For each partition size with index *i*, the CTU level estimation is only executed in the case, where the maximum coefficient C_{max} is greater than a threshold τ_1 and the number of non-zero coefficients I_{Nz} is greater than a threshold τ_2. This is motivated by the fact that in CTUs where most coefficients are of very low magnitude, the quantization step cannot be identified due to noise. An estimation of the frame level QP is always calculated.

Motivated by the CTU level QP constraint in the rate control scheme as described in Section 2.2, once the CTU level QP estimation matrix $[\widehat{QP}]^i_{M\times N}$ is obtained, each QP estimate at (m,n), $0 \leqslant m \leqslant M,\ 0 \leqslant n \leqslant N$ will be evaluated using the following constraint to rule out the outliers due to initial local decisions:

$$QP_{currLCU\pm L} - Offset \leqslant QP_{currLCU} \leqslant QP_{currLCU\pm L} + Offset \qquad (16)$$

where *L* is the length of a 1-D sliding window and *offset* denotes the acceptable QP variations within the whole frame. Only the QP estimates satisfying (16) in matrix $[\widehat{QP}]^i_{M\times N}$ are considered as valid for further analysis. The improvements given by (16) for QP estimation are reported in Section 6.1.

Let N^i_{tot} be the total number of CTUs with valid QP estimations, i.e. satisfying the constraint imposed by the thresholds τ_1, τ_2 and (16), in a residual frame with index *i*. Only if the proportion of CTUs with valid QP over the whole frame is greater than the threshold τ_3, this partition scenario with index *i* will be further processed. Once the valid matrix $[\widehat{QP}]^i_{M\times N}$ and frame level QP estimation $\widehat{QP}^i_{frame}$ is obtained, a decision is made to choose between the CTU level and the frame level QP estimation by using a threshold τ_4. In either case, we count how many CTUs have the corresponding QP estimation values (in a range of $QP \pm \Delta QP$, where $\pm\Delta QP$ guarantee only the best estimation values in this partition are accumulated) and save these counts in an array. By accumulating the counter array with four different partitions, the final estimation QP_{est} for the intra frame is assigned to the QP value with the largest counter value. The procedure is described in detail in Algorithm 1.

In this algorithm QP^i_{est} denotes the estimated QP for residual frame *i*. $[C]^i_{1\times 32}$ is an array to save the occurrence counts of the corresponding QP elements, where $\mathrm{QP} \in [20,\ldots,51]$. $N_{max}([\widehat{QP}]^i_{M\times N})$ is a function that returns the QP value appearing the most times as the QP value in matrix $[\widehat{QP}]^i_{M\times N}$. $\theta(QP)$ provides the count of corresponding QP value at the estimation matrix $[\widehat{QP}]^i_{M\times N}$. $N_{frame}(= M \times N)$ is the total number of CTUs in the frame. In our implementation, the parameters are set to $\tau_1 = 100$, $\tau_2 = 25$, $\tau_3 = 0.25$ and $\tau_4 = 3$. *L* and *offset* are set to 5 and 10, respectively. ΔQP is set to 3. We evaluate the performance of the described QP

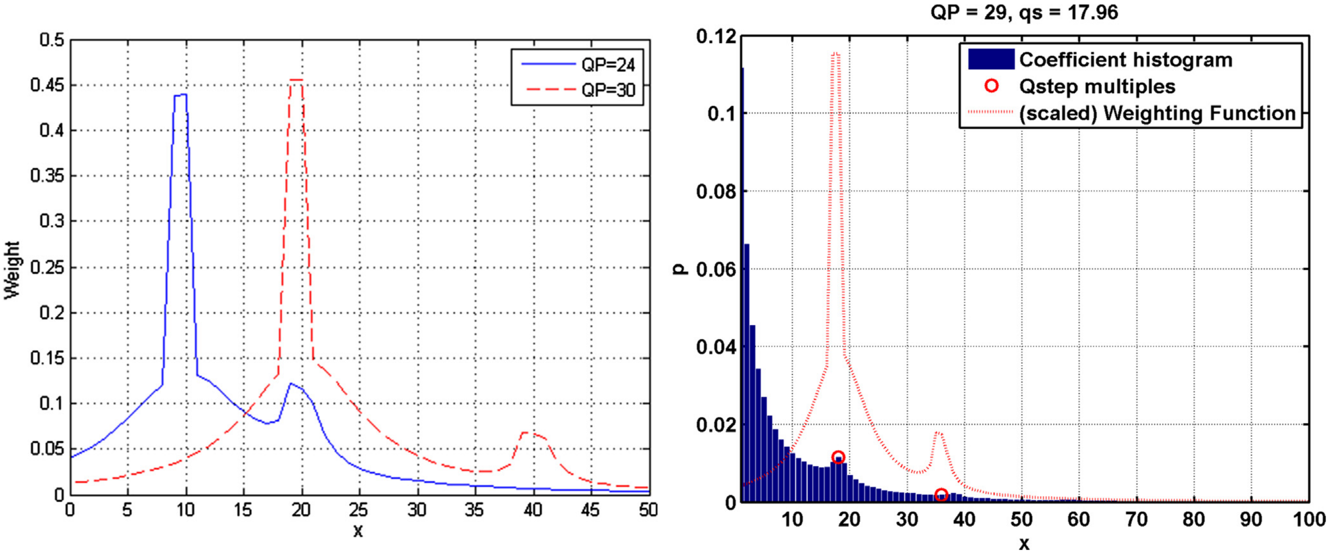


Fig. 3. Left: Examples of the weighting function $\hat{f}(x, QP)$ from (12) for residuals with partition size 4 × 4 and $QP = 24$ and 30, respectively. Right: Histogram of all transform coefficients belonging to a TU of size 4 × 4 and with QP = 27 (for which $qs(QP) = 17.96$) in the LIVE training videos.

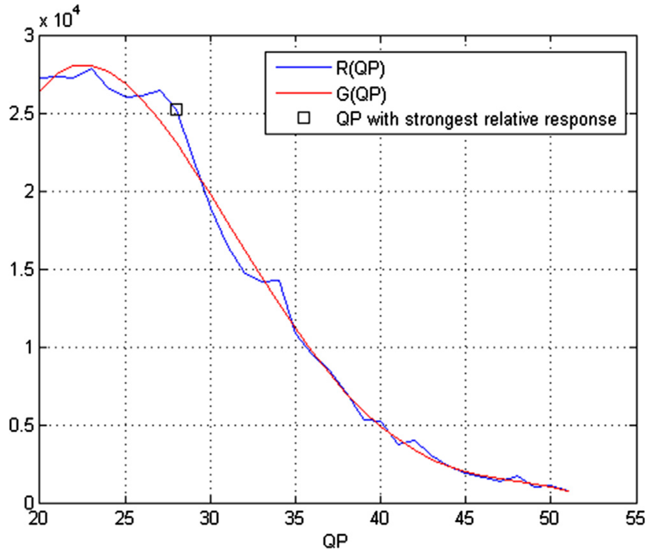


Fig. 4. Response function $R(QP)$ and fitting polynomial $G(QP)$ function. The value of estimated QP (= 28, *true* QP = 28) with the strongest relative response is marked by a black-square. The data is from a typical CTU from the first I-frame in an HEVC encoded video at 1024 kbit/s of the "Blue Sky" sequence from the LIVE dataset.

estimation scheme both on a training set and a test set. We report the results in Section 6.

Algorithm 1. QP estimation scheme.

```
Initial counter array $[C]^i_{1\times 32}$ for all possible $QP \in [20,\ldots,51]$
  values with zeros
for $i \leftarrow 1–4$
  for each CTU at $(m,n)$
    if $C_{max} \geqslant \tau_1 \&\& I_{Nz} \geqslant \tau_2$
      calculate $\widehat{QP}^i_{(m,n)}$ based on (15)
    end
  end
  calculate $\widehat{QP}^i_{frame}$ based on (15)
  processing $[\widehat{\mathrm{QP}}]^{\mathrm{i}}_{\mathrm{M\times N}}$ based on (16)
  if $N^i_{tot}/N_{frame} \geqslant \tau_3$
    calculate $QP_{CTU} = N_{max}([\widehat{QP}]^i_{M\times N})$
    if $\theta(QP_{CTU})/\theta(\widehat{QP}^i_{frame}) \geqslant \tau_4$
      $QP^i_{est} = QP_{CTU}$
    else
      $QP^i_{est} = \widehat{QP}^i_{frame}$
    end
    for $qp_{idx} \leftarrow QP^i_{est} - \Delta QP$ to $QP^i_{est} + \Delta QP$
      update $[C]^i_{1\times 32}$ with $\theta\,(qp)$
    end
  end
end
calculate $QP_{est} = argmax_{QP} \sum_{i=1}^{4} [C]^i_{1\times 32}$
```

To build features that are mapped to a video quality score using machine learning in Section 5.2, besides the overall estimated QP for the frame, the following statistics for the corresponding residual frame are collected for our analysis:

– QP^i_{est} (the estimated QP for residual frame i) in Algorithm 1.
– P^i_{con} (the fraction N^i_{QP}/N^i_{tot} which is used as a measure of confidence. N^i_{QP} denotes the number of CTUs with valid estimated QP^i_{est})
– P^i_{tot} (the fraction N^i_{tot}/N^i_{frame}).

4.1.1. Computational complexity

The computational complexity of the proposed method may briefly be compared and related to the complexity of a HEVC intra encoder. For codec features, we do not use any temporal analysis besides pooling in our method, since our analysis is performed frame by frame.

Besides the codec features, we also use the spatial perceptual information measure (SI) and the temporal perceptual information measure (TI) [2] of each video sequence as complementary features as described in Section 5.1.

SI/TI information is simpler to calculate even than most IQA methods since the temporal aspect is only utilized in TI, which is solely based on consecutive frame differences. The most computationally complex aspect of our method is the mimicking of the intra-prediction in HEVC, which is simplified since we only use 4 different CU candidate sizes. However, (15) is calculated several times during the estimation of the QP values in each CTU, which adds to the complexity.

4.2. GOP estimation

Since the proposed algorithm is designed to estimate QP of HEVC Intra coded frames, the positions of the Intra frames have to be detected (if they are not known a priori). In this work, fixed GOP size for the video sequences is assumed. If the GOP size varies, alternative methods such as [32] are needed to estimate the positions of the Intra frames. With fixed GOP, the positions of the Intra frames can be estimated from the decoded video based on the P^i_{con} statistics obtained in Section 4.1.

As a measure of reliability a vector $\boldsymbol{c}$ is produced based on P^i_{con}, where the value of an element c_j, represents the numerical confidence of the analysis for the corresponding frame:

$$c_j = \sum_{i\in[1,4]} P^i_{con}(j) \tag{17}$$

where $i \in [1,4]$ is an index of the prediction size from 4×4 to 32×32 and j is the index of the frame number. In order to reduce the dependency of the elements in $\boldsymbol{c}$ on the content, we subtract the mean and the standard deviation of the vector from the values

$$\tilde{c}_j = c_j - mean(\boldsymbol{c}) - std(\boldsymbol{c}). \tag{18}$$

Then, for each candidate GOP size s, filter vector $\boldsymbol{v}(s)$ is created, where the elements at positions equal to $ks+1$ for all $k \in Z+$ are set to 1 and all other elements are set to 0. Then, we calculate a response denoted $R_c(s)$ by summing the element-wise product of $\tilde{\boldsymbol{c}}$ and the filter vector,

$$R_c(s) = \begin{cases} \sum_{j=1}^{N} \tilde{c}_j v_j(s) & if\ \frac{N_p(\tilde{\boldsymbol{c}}\odot\boldsymbol{v}(s))}{N_{non}(\tilde{\boldsymbol{c}}\odot\boldsymbol{v}(s))} > T \\ 0 & otherwise \end{cases} \tag{19}$$

where j is the index of the elements in the vectors, N is the number of analyzed frames, $\tilde{\boldsymbol{c}} \odot \boldsymbol{v}(s)$ denotes the element-wise product of $\tilde{\boldsymbol{c}}$ and $\boldsymbol{v}(s)$, $N_p(\tilde{\boldsymbol{c}} \odot \boldsymbol{v}(s))$ returns the number of positive elements in the vector, and $N_{non}(\tilde{\boldsymbol{c}} \odot \boldsymbol{v}(s))$ returns the number of non-zero elements in the vector. T is set to 0.3. The candidate GOP size s with the biggest response value of $R_c(s)$ is selected as the estimation of the GOP size.

In the expression above, we assume that the start of the analyzed video is also the beginning of a GOP, i.e. the analyzed video starts with an Intra frame. We can relax this assumption by introducing an integer offset j, so that every element at positions equal to $ks + j$ for $k \in Z+$ is set to 1 in the filter vector. In our experiments, candidate GOP sizes in the interval $[1, 2, \ldots, 100]$ are used. The GOP estimation is perfectly accurate over 40 HEVC coded sequences from the LIVE dataset when the offset parameter for the filter vector is not used. Without the assumption that the analyzed video starts with an Intra frame, only for 1 out of the 40 videos the GOP is wrongly estimated. Due to the generally high accuracy, in the further experiments we assume that the position of the Intra frames is known, when we are estimating the quality. Examples of GOP estimation on LIVE and SJTU sequences are shown in Fig. 5.

4.3. PSNR estimation

In this subsection, we present a brief description of a PSNR estimation method for I-frames in HEVC videos based on the codec analysis. It adapts the approach for H.264/AVC described in [18], wherein more details are described. It should be noted that the PSNR estimate is not the final goal in this work, but it is considered and may be used as an intermediate result and also for comparison to the related work in [21]. As detailed in [33], the local mean squared error ε_k^2 of a block at the kth coefficient band can be approximated by using the quantized value X_k

$$\varepsilon_k^2(block) \approx \frac{\int_{a_k}^{b_k} f_X(x)(X_k - x)^2 dx}{\int_{a_k}^{b_k} f_X(x)dx} \tag{20}$$

where $f_X(x)$ is the original distribution of the coefficients at band k.

We use a Cauchy distribution (7) for each band k motivated by [30] to model the original distributions of the transform coefficients. The shape parameter of each Cauchy distribution at band k can be estimated by the percentage of coefficients smaller than $\alpha \cdot qs$ in the reconstructed coefficients as in [34]. With a given Cauchy distribution, we are able to calculate an approximation for the local mean squared error. Different from the work on H.264/AVC [18], where variations of transform block size are considered in the PSNR estimation, we only consider the analysis from the transform block size of 4 × 4 in the PSNR estimation for the sake of simplicity. By using (20), the mean squared error is calculated as in [18] and finally, the estimation of the PSNR over the whole frame is obtained by accumulating the errors of 16 different bands of 4 × 4 block.

$$PSNR_{esti} = 10log_{10}\frac{255^2}{\left(\sum_{k=1}^{16}\hat{\varepsilon}_{k,j}^2\right)/16}. \tag{21}$$

5. Video features and machine learning

In order to obtain a mapping from the feature space of the video analysis to a quality score we use machine learning. Different machine learning methods have been used for image and video quality assessment, e.g. [35–40]. To keep complexity low and avoid the "black box" issue of some machine learning methods, we have chosen to use the relatively simple Elastic Net method in this work. The Elastic Net is a sparse linear regression method with shrinkage. The method is outlined in Section 5.2 while details can be found in [41,42]. The codec analysis features are generated based on the analyzed results described in Section 4.1. The estimated PSNR in Section 4.3 is not chosen for feature generation in this paper, since it is highly correlated with the estimated QP and it does not improve the performance of our subjective quality assessment based on our experiments. In addition to the codec features, SI/TI values are introduced as complementary features.

5.1. Feature generation

We use the average and standard deviation of the analyzed results (Section 4.1) from Intra frames as basic temporal poolings for the HEVC codec features. We perform the pooling over the full length of the sequences. Besides the best QP estimation QP_{est} (Algorithm 1), QP_{est}^i on four different residual frames i (one for each partition as discussed in Section 2), P_{con}^i and P_{tot}^i introduced in Section 4.1, a weighted QP estimate wQP for each Intra frame indexed by k is calculated based on QP_{est}^i as:

$$wQP(k) = \sum_i QP_{est}^i(k)P_{con}^i(k) / \sum_i P_{con}^i(k). \tag{22}$$

The features generated for mapping to a quality score are: the sum, the mean, and the maximum of P_{con}^i and P_{tot}^i, the mean, the standard deviation, the maximum differences of wQP, the gradient of a straight line fitted to wQP and the maximum difference between two consecutive I-frames of wQP. Additionally, the mean, the standard deviation, the maximum differences of QP_{est}, the gradient of a straight line fitted to QP_{est}, and the maximum difference between two consecutive I-frames of QP_{est} are also selected as features. Similar to the clustering used in [18], the best QP estimation values are grouped into two clusters, the mean of good and bad quality of the QP_{est} and a weighted mean of QP_{est} is calculated. Thus a total of 20 features are initially calculated based on the HEVC analysis as codec analysis.

To get information about the spatial and temporal complexity in the videos we calculate the SI and TI measures from [2] on the distorted videos. Since we calculate the SI/TI values from the distorted videos, they will depend on the amount of distortion as shown in [43]. Nevertheless, the measures still contain information about the spatial and temporal complexity of the videos, which will be useful in our machine learning approach. Instead of using the maximum of the spatial and temporal complexity values over time as in [2], we instead use the average and standard deviation of the SI/TI values of each frame as features, resulting in a total of 4 additional features. This has also previously been used e.g. in [44].

The relation between spatial/temporal complexity and perceptual quality at different target bitrates for HEVC videos was also studied in [45]. The complexity was not measured as recommended in [2], but for high quality HEVC videos it was found that spatial complexity generally reduced the subjective score, while temporal complexity generally increased the subjective score.

5.2. Elastic net

To map the features from the video codec analysis we use the Elastic Net as implemented in [46]. The goal of the method is to estimate the coefficients β of a regularized linear regression model:

$$\tilde{\beta} = \underset{\beta}{argmin}\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \tag{23}$$

where y represents the target quality values, X is a feature matrix, and λ_1 and λ_2 are regularization parameters of the $L1$-norm and the $L2$-norm, respectively. The optimal values of the regularization parameters are found using cross-validation.

Due to the $L1$-norm in (23) the solution of an Elastic Net can generally be considered to be sparse. Even so, since our training set only consists of a relatively small number of different original video contents, we perform feature selection before applying the Elastic Net as described above in order to limit the risk of overfit-

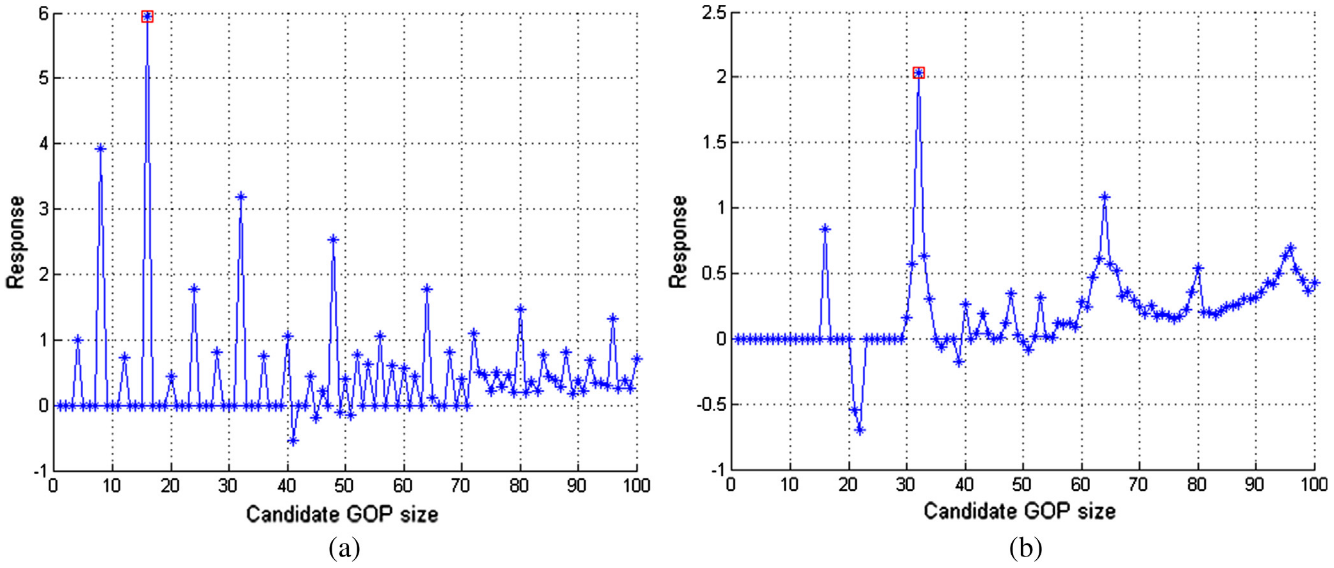


Fig. 5. Examples of the response function *Rc* as function of candidate GOP sizes when using the offset. (a) "Pedestrian Area" video coded at 1024 kbit/s from the LIVE database. (b) "Runners" video coded at 6336 kbit/s from SJTU database. The value of the true GOP size is marked by a red-square. This value coincides with the maximum response value of *Rc*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ting. The feature selection is inspired by [47] where n artificial noise features are introduced and used in the training of a modified Support Vector Machine (SVM). Instead, we use the Elastic Net with artificial noise features. Since the absolute output $|\widetilde{\beta}|$ of the Elastic Net can be considered as weights of the features, this can be used for ranking of the features and to deselect features with a weight lower than that of the average of the weights of the noisy features. In our implementation, we sum the weights over the cross-validation folds in the training set with $n = 3$. This approach is repeated until no features are deselected anymore and the remaining subset of features then constitutes the set of features chosen which is used to train the regression model as explained above.

Based on our experiment, 9 out of 24 features are deselected for the cross-validation tests and 5 features are deselected for the cross dataset tests described in Sections 6.2 and 6.4, respectively. Using the absolute and normalized $\tilde{\beta}$-values, denoted $|\tilde{\beta}_N|$, we can find the most influential features of the model and the features selected. When based on the feature selection cross-validation output, we found the values of $|\tilde{\beta}_N|$ to be between 0.07 and 0.16. In this case, the 4 features with highest $|\tilde{\beta}_N|$ values and accounting for more than half of the sum were in descending $|\tilde{\beta}_N|$ order: the standard deviation of *wQP* (22), the maximum difference between two consecutive I-frames of QP_{est} from Algorithm 1, the maximum of P^i_{con} from Algorithm 1, and the gradient of a line fitted to *wQP* (22). These four features can be thought of as representing the global variation of the quantization, the highest local variation of the quantization, the confidence of the estimation (which is related to the content), and the overall tendency of the quantization (increasing or decreasing), respectively. Regarding the SI/TI values, the minimum value of the TI was found to have a negative β-value in the model, indicating a negative influence on the perceived quality, while the SI-values were excluded from the final model.

6. Experimental results

In order to fairly evaluate the proposed scheme, video sequences from two different datasets, LIVE [31] and SJTU [48], are used. We encode all 10 reference sequences from the LIVE dataset with resolution 768 × 432 with HEVC at 4 bitrate levels, and consider these coded sequences (in total 40 sequences from LIVE dataset) as a training set for the coefficient modeling and QP/PSNR estimation scheme. We refer to this HEVC training set as the LIVE sequences. For verification of the method, we use the SJTU dataset for testing, which contains sequences with a resolution of 3840 × 2160. In this dataset, 10 different sequences are HEVC coded at 6 bitrate levels.[1] Thus, a total of 60 sequences with opinion scores are available.

All sequences, i.e. both LIVE and SJTU, are compressed with HEVC reference software (HM 11.0) using the standard rate control algorithm, thus the distortion is solely due to compression. The characteristics of the datasets are summarized in Table 2. To make it clear, the LIVE sequences encoded by us are used in this work for training and modeling purposes, while the publicly available and unmodified SJTU dataset is used for cross-validation and independent verification.

The chosen sequences contain a wide variety of content, from natural sequences to sequences generated digitally, from low motion sequences to intensive motion sequences. Fig. 6 depicts SI and TI calculated for the original sequences in the LIVE and SJTU datasets according to P.910 recommendation [2]. For the vertical axis, low values correspond to sequences having limited motion, and high values indicate that a sequence contains scenes with intensive motion. For the horizontal axis, low values correspond to scenes having minimal spatial detail, and high values are found in scenes having significant spatial detail.

6.1. Performance of QP and PSNR estimation

Based on the test conditions above, the QP and PSNR estimation described in Sections 4.1 and 4.3 are evaluated. Root Mean Square Error (RMSE) is used to measure the difference between estimated QP/PSNR and the true QP/PSNR in both training and verification datasets. The Spearman Rank Order Correlation Coefficients (SROCC) and the Linear Correlation Coefficients (LCC) between estimated PSNR and the true PSNR are calculated. Estimated vs. true

[1] The specific bitrates can be found at http://medialab.sjtu.edu.cn/resources/SubScore2.pdf.

Table 2
The characteristics of the video sequences.

	LIVE	SJTU
Original videos	10	10
Bitrate levels	4	6
Resolution	768 × 432	3840 × 2160
Bitrates [Kbps]	256–1024	730–24,000
Framerate [fps]	25–50	30
Duration [s]	10	8–10
GOP size	16	32
GOP structure	IBBBBB	IBBBPBBBP

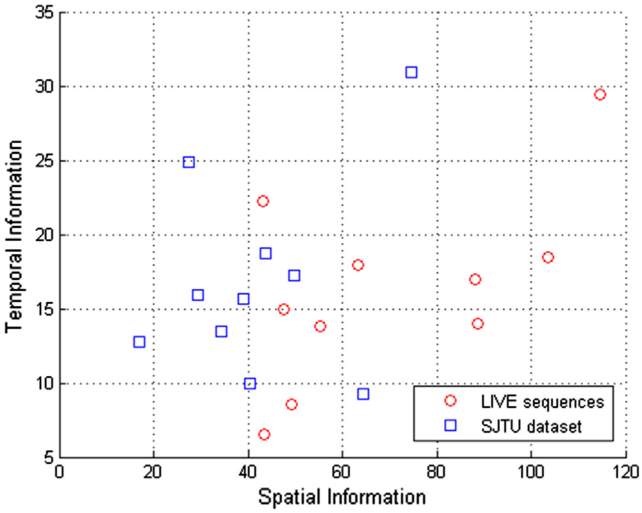


Fig. 6. Scatterplot (SI,TI) of the spatial-temporal information for the original video sequences from the LIVE and SJTU datasets.

QP values are shown in Fig. 7, the RMSE is equal to 2.85 for the LIVE training sequences and 2.90 for the SJTU dataset for the QP estimation. To show the improvement given by (16), the QP estimations without (16) are also reported here, in which case the RMSE is equal to 3.12 for the LIVE training sequences and equal to 3.48 for the SJTU dataset.

The performance of our PB PSNR estimation is equal to a SROCC of 0.94, LCC of 0.90 and RMSE of 3.5 for the LIVE training sequences. For the SJTU dataset, the SROCC is equal to 0.86, the LCC is equal to 0.82 and RMSE is equal to 2.6. For PSNR estimation, we compare to a NR BB VQA method for HEVC presented in [21], in which high accuracy PSNR estimates with correlations equal to 0.97–0.98 are reported. As NR PB VQA, the proposed method in this paper does not access the bitstream but can still achieve relatively high correlation with objective PSNR measurements. As stated above, neither QP nor PSNR estimation is the final goal of our analysis, but rather to build useful features for the mapping to subjective video quality scores.

6.2. Performance of proposed VQA

Based on the generated features using an Elastic Net as described in Section 5, we are able to map those features to a subjective quality score. The performance of the proposed VQA scheme is evaluated in this section. In all cross-validation test cases, we used all sequences in the SJTU dataset by leaving out 2 video contents of the total 10 video contents for testing in each cross-validation fold. The remaining videos were used as the training set for that fold of the cross-validation. The experiments were done for all possible content-independent splits between training and test data. This test procedure results in 45 splits. When using content-independent splits all videos that have been coded using the same original video is either in the training or in the test set of a cross-validation fold and never split in any way between the two sets. The median $\tilde{x}$, mean μ, and standard deviation σ of the SROCC, of the LCC, and of the RMSE for the proposed NR PB VQA are given in Table 3.

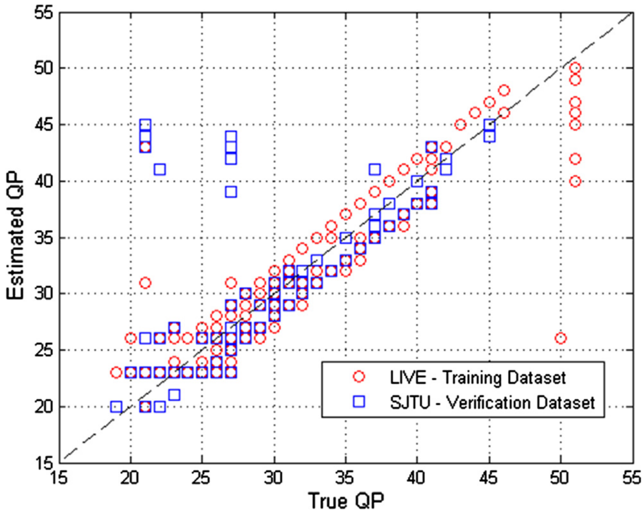


Fig. 7. HEVC no-reference QP estimation vs. true QP for intra frames.

As shown in Table 3, in the cases where only codec analysis features are used, good results for predicting the video quality scores are achieved, with a mean SROCC of 0.89, a mean LCC of 0.86, and a mean RMSE of 0.32. Adding SI/TI information slightly improves the performance of the proposed method, with a mean SROCC of 0.90, a mean LCC of 0.87, and a mean RMSE of 0.29. To the best of our knowledge, these are the first NR PB VQA results for HEVC UHD sequences.

For comparison, Table 4 lists results for the case where the true QPs (QP_{true}) are utilized rather than the estimates from Section 4.1 as also reported in [24]. This may also be seen as a NR VQA metric with bitstream access, i.e. a NR BB VQA scheme. Since the QP_{true} in the bitstream is parsed (just reading the frame level QP), less features are calculated in the codec analysis for machine learning compared to our NR PB VQA method. A total of 9 QP_{true} relevant features are calculated in NR BB VQA. The 5 first features includes: the mean, the standard deviation, the maximum of the differences of QP_{true}, the gradient of a straight line fitted to QP_{true}, and the maximum difference between two consecutive I-frames of QP_{true}. By clustering the QP_{true} into two clusters we calculate the last 4 features as the mean of good and bad quality of the QP_{true}, the weighted mean of QP_{true} in the clusters and the weight that is based on the ratio of the means in the clusters. These features are similar to the features generated in the PB case, but fewer features are needed as there is no uncertainty regarding the QP values. The results show potential for improvements of NR PB VQA if better QP estimation is developed. If the bitstream is accessible, the NR BB VQA gives better results, and will also be less complex as our parsing of the bitstream for I frame QP values is very simple and thereby faster.

6.3. Comparison to classic and state-of-the-art metrics

To illustrate the competitive performance of the proposed method, we test the performance of well-known FR metrics, such as PSNR, SSIM [23], MS-SSIM [23], VQM [49], VIF [50], and STMAD [51], with the same cross-validation procedure as in Section 6.2 for fair comparison. That is, 2 video contents out of the total 10 video

Table 3
Cross-validation results on SJTU dataset (NR PB VQA).

	SROCC			LCC			RMSE		
	$\tilde{x}$	μ	σ	$\tilde{x}$	μ	σ	$\tilde{x}$	μ	σ
Codec features	0.90	0.89	0.09	0.87	0.86	0.08	0.32	0.32	0.08
Codec features + SI&TI	0.92	0.90	0.07	0.88	0.87	0.07	0.29	0.29	0.07

Table 4
Cross-validation results on SJTU dataset by NR BB VQA (using true QP from HEVC decoder).[24]

	SROCC			LCC			RMSE		
	$\tilde{x}$	μ	σ	$\tilde{x}$	μ	σ	$\tilde{x}$	μ	σ
Codec features	0.94	0.93	0.05	0.91	0.90	0.04	0.27	0.27	0.05
Codec features + SI&TI	0.95	0.94	0.04	0.92	0.91	0.04	0.25	0.25	0.06

Table 5
Mean SROCC and LCC for FR metrics on SJTU dataset with the same cross-validation procedure.

	PSNR	SSIM	MS-SSIM	VQM	VIF	STMAD	Proposed
SROCC	0.82	0.84	0.92	0.85	0.93	0.87	0.90
LCC	0.77	0.78	0.88	0.87	0.87	0.86	0.87

contents in SJTU dataset are tested for FR metrics in each cross-validation fold, which results in 45 splits. The mean values of the SROCC and LCC values between FR metrics and MOS are given in Table 5. Note that these FR metrics are not only intended for video distortions caused by the encoding and they are in that sense more general than the proposed method. The newly developed image metric VSI [52] was also tested, but with simple pooling it showed poor performance on our video datasets and was therefore not included in the comparison.

As it appears from Table 5, MS-SSIM and VIF give better correlations with subjective video quality scores compared to the other FR metrics on SJTU dataset, which is also consistent with the results in [16]. Comparison to the performance of the proposed NR PB VQA scheme is also reported in Table 5, showing that the proposed scheme provides better performance than the FR metrics PSNR, SSIM, VQM, and STMAD. Meanwhile, it gives comparable performances compared to FR metrics MS-SSIM and VIF. Comparison to the performance of the NR BB VQA scheme as reported in Table 4, shows that the proposed scheme using bitstream access is superior to the performance of all the FR metrics listed in Table 5.

For further analysis, we also applied a statistical analysis of the SROCCs as shown in Fig. 8. Using multi-comparison ANOVA with a confidence level of 95% on the SROCC performance of our proposed NR PB VQA method (as well as NR BB VQA) with codec feature and SI/TI features versus the FR in the SJTU dataset revealed that NR PB VQA, NR BB VQA, and the FR methods MS-SSIM and VIF have statistically similar performance and they are all significantly better than PSNR. NR BB VQA, MS-SSIM and VIF are furthermore statistically better than SSIM and VQM.

For the sake of completeness, the SROCC and LCC between FR metrics and MOS over all test sequences in the SJTU dataset rather than using the cross-validation procedure are given in Table 6. Note that the performance of the proposed method in Table 6 is the cross-dataset performance as presented in Section 6.4, which is limited by the differences in the two datasets that is also discussed in that Section.

We have also considered comparisons to NR methods. The distortion-generic NR IQA method BRISQUE [37] was computed for the videos in the SJTU dataset. The LCC and SROCC for the mean BRISQUE scores over the frames of a video on this dataset were 0.33 and 0.32, respectively. A NR BB method for HEVC predicting FR VQA scores was presented in [38]. In the leave-one-out cross-validation test reported for predicting VQM scores they obtained a mean LCC, SROCC, and RMSE of respectively 0.98, 0.98, and 0.11. However, this result is without predicting subjective scores and with only 1 video content used for testing in each fold. The NR VQA method Video BLIINDS [39] was also considered for comparison, but due to the complexity of the method and the data size of the UHD videos in the SJTU dataset, it was infeasible to perform the computational calculations. However, in [39] it is stated that Video BLIINDS does not quite attain the performance level of the FR method STMAD, which is used for comparison in this work.

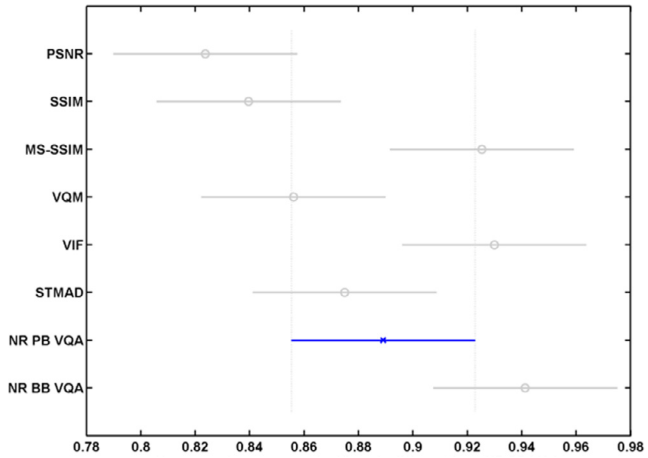


Fig. 8. Multi-comparison ANOVA with a confidence level of 95% on the cross-validation SROCC performance.

6.4. Cross-dataset verification

As we can see from Tables 5 and 6, among the FR methods, MS-SSIM and VIF give better correlations with subjective video quality scores. As verification of the robustness of the proposed method and the independence of a particular test dataset, we train the model on one dataset using the FR method, MS-SSIM, and test on the other independent dataset. Thus, we train our algorithm on

Table 6
SROCC and LCC for FR metrics measured on all the SJTU dataset sequences.

	PSNR	SSIM	MS-SSIM	VQM	VIF	STMAD	Proposed
SROCC	0.73	0.75	0.90	0.78	0.91	0.80	0.83
LCC	0.67	0.72	0.84	0.79	0.85	0.77	0.77

Table 7
SROCC and LCC for cross-dataset validation.

	SROCC	LCC
LIVE:MS-SSIM SJTU:MOS	0.83	0.77
LIVE:MS-SSIM SJTU:MS-SSIM	0.81	0.78

all 40 HEVC video sequences using the LIVE training sequences with MS-SSIM as the subjective video quality assessment metric. Thereafter, we tested this model on all 60 HEVC 4 K videos in the SJTU dataset with either MOS or MS-SSIM as the video quality assessment metric. There is no overlap in videos between the two datasets, so this split is also content-independent. Additionally, the two datasets are very different from each other, e.g. the video resolution is equal to 768 × 432 in the LIVE dataset, while 4 K UHD is used in the SJTU dataset.

For these cross-dataset settings, we still achieve promising results as reported in Table 7 with a SROCC equal to 0.83 and a LCC equal to 0.77 if we take MOS as the quality descriptor for the SJTU dataset. If we take MS-SSIM as the quality descriptor for the SJTU dataset, we achieve a SROCC equal to 0.81 and a LCC equal to 0.78. Comparison to the performance of FR metrics is reported in Table 6, showing that the proposed scheme provides better performance than PSNR and SSIM in this test. One would expect this performance to improve if the model took the differences between the datasets into account, e.g. the video resolution, but this is left for future work.

7. Conclusion

In this paper, a framework for a NR PB VQA method for HEVC videos based on codec analysis was presented. The proposed method is only based on the decoded video without accessing the bitstream. This is achieved by codec video analysis for which a novel generation of the residuals after intra prediction based on the HEVC quad-tree structure, a coefficient model for the new transform and filter introduced in HEVC, and a joint frame and Coding Tree Unit (CTU) level analysis scheme for QP estimation were presented. The proposed method is able to estimate the QP used in the Intra frames and to generate features for VQA purposes based on the codec analysis and spatio-temporal complexity information. These features were mapped to a subjective quality score using an Elastic Net. Extensive testing was performed on a HEVC coded 4 K UHD video quality dataset. By combining generated codec features and SI/TI information in the NR VQA we achieved a median SROCC value of 0.92 for the SJTU HEVC 4 K dataset. The results showed that the proposed scheme achieved robust performance on the HEVC test videos and that the estimated quality scores are highly correlated with the subjective quality scores. Furthermore, even compared to the state-of-the-art FR VQA metrics, the proposed method shows competitive performance.

Acknowledgment

This work was supported in part by SVCF Grant 117983 from the Cisco University Research Program Fund.

References

[1] G.J. Sullivan, J. Ohm, W.J. Han, T. Wiegand, Overview of the High Efficiency Video Coding (HEVC) standard, IEEE Trans. Circ. Syst. Video Technol. 22 (12) (2012) 1649–1668.
[2] International Telecommunication Union (ITU), ITU-T Rec. P.910—Subjective Video Quality Assessment Methods for Multimedia Applications, 1999.
[3] S. Winkler, A perceptual distortion metric for digital color video, Proc. SPIE Human Vi0073ion Electron. Imag. IV, vol. 3644, 1999, pp. 175–184.
[4] T. Oelbaum, K. Diepold, A reduced reference video quality metric for AVC/H.264, in: Proc. EUSIPCO, 2007, pp. 1265–1269.
[5] W. Xue, X. Mou, L. Zhang, A.C. Bovik, X. Feng, Blind image quality assessment using joint statistics of gradient magnitude and laplacian features, IEEE Trans. Image Process. 23 (11) (2014) 4850–4862.
[6] A. Mittal, R. Soundararajan, A.C. Bovik, Making a completely blind image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212.
[7] K. Gu, G. Zhai, X. Yang, W. Zhang, Hybrid no-reference quality metric for singly and multiply distorted images, IEEE Trans. Broadcast. 60 (3) (2014) 555–567.
[8] Q. Jiang, F. Shao, G. Jiang, M. Yu, Z. Peng, Supervised dictionary learning for blind image quality assessment using quality-constraint sparse coding, J. Vis. Commun. Image Represent. 33 (2015) 123–133.
[9] Q. Wu, H. Li, F. Meng, K.N. Ngan, B. Luo, C. Huang, B. Zing, Blind image quality assessment based on multichannel feature fusion and label transfer, IEEE Trans. Circ. Syst. Video Technol. 26 (3) (2016) 425–440.
[10] Q. Wu, H. Li, F. Meng, K.N. Ngan, S. Zhu, No reference image quality assessment metric via multi-domain structural information and piecewise regression, J. Vis. Commun. Image Represent. 32 (2015) 205–216.
[11] Q. Wu, Z. Wang, H. Li, A highly efficient method for blind image quality assessment, in: Proc. ICIP, 2015, pp. 339–343.
[12] M. Shahid, A. Rossholm, B. Lövström, H.J. Zepernick, No-reference image and video quality assessment: a classification and review of recent approaches, EURASIP J. Image Video Process. 1 (2014) 1–32.
[13] J. Søgaard, S. Forchhammer, J. Korhonen, No-reference video quality assessment using MPEG analysis, in: Picture Coding Symposium, San Jose, 2013, pp. 161–164.
[14] T. Na, M. Kim, A novel no-reference PSNR estimation method with regard to de-blocking filtering effect in H.264/AVC bitstreams, IEEE Trans. Circ. Syst. Video Technol. 24 (2) (2014) 320–330.
[15] G. Valenzise, M. Tagliasacchi, S. Tubaro, Estimating QP and motion vectors in H.264/AVC video from decoded pixels, in: Proc. 2nd ACM Workshop Multimedia Forensics, Security, Intelligence, New York, 2010, pp. 89–92.
[16] K. Zhu, C. Li, V. Asari, D. Saupe, No-reference video quality assessment based on artifact measurement and statistical analysis, IEEE Trans. Circ. Syst. Video Technol. 25 (4) (2015) 533–546.
[17] S. Forchhammer, H. Li, J.D. Andersen, No-reference analysis of decoded MPEG images for PSNR estimation and post-processing, J. Vis. Commun. Image Represent. 22 (4) (2011) 313–324.
[18] J. Søgaard, S. Forchhammer, J. Korhonen, No-reference video quality assessment using codec analysis, IEEE Trans. Circ. Syst. Video Technol. 25 (10) (2015) 1637–1650.
[19] Y. Li, L.-M. Po, C.-H. Cheung, X. Xu, L. Feng, F. Yuan, K.-W. Cheung, No-reference video quality assessment with 3D shearlet transform and convolutional neural networks, IEEE Trans. Circ. Syst. Video Technol. 26 (6) (2016) 1044–1057.
[20] Y. Xue, B. Erkin, Y. Wang, A novel no-reference video quality metric for evaluating temporal jerkiness due to frame freezing, IEEE Trans. Multimedia 17 (1) (2015) 134–139.
[21] B. Lee, M. Kim, No-reference PSNR estimation for HEVC encoded video, IEEE Trans. Broadcast. 59 (1) (2013) 20–27.
[22] T. Lin, N. Yang, R. Syu, C. Liao, W. Tsai, C. Chou, S. Chen, NR-Bitstream video quality metrics for SSIM using encoding decisions in AVC and HEVC coded videos, J. Vis. Commun. Image Represent. 32 (2015) 257–271.
[23] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[24] X. Huang, J. Søgaard, S. Forchhammer, No-reference video quality assessment by HEVC codec analysis, in: IEEE Int'l Conf. Visual Comm. Image Proc. (VCIP), 2015.
[25] G.V. Wallendael, N. Staelens, L. Janowski, J. Cock, P. Demeester, R.V. Walle, No-reference bitstream-based impairment detection for high efficiency video coding, in: IEEE Int'l Workshop on Quality of Multimedia Experience, 2012.
[26] J. Lainema, F. Bossen, W. Han, J. Min, K. Ugur, Intra coding of the HEVC standard, IEEE Trans. Circuits Syst. Video Technol. 22 (12) (2012) 1792–1801.
[27] B. Li, H. Li, L. Li, J. Zhang, Rate Control by R-Lambda Model for HEVC, ITU-T SG16 Contribution, JCTVC-K0103, Shanghai, October 2012.
[28] M. Budagavi, A. Fuldseth, G. Bjontegaard, V. Sze, M. Sadafale, Core transform design in the High Efficiency Video Coding (HEVC) standard, IEEE J. Sel. Top. Signal Process. 7 (2013) 1029–1041.

[29] C.M. Fu, E. Alshina, A. Alshin, Y.W. Huang, C.Y. Chen, C.Y. Tsai, W.J. Han, Sample adaptive offset in the HEVC standard, IEEE Trans. Circ. Syst. Video Technol. 22 (12) (2012) 1755–1764.

[30] Y. Altunbasak, N. Kamaci, An analysis of the DCT coefficient distribution with the H.264 video coder, in: IEEE Int'l Conf. Acoust. Speech Signal Process., Montreal, 2004, pp. 177–180.

[31] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, L.K. Cormack, Study of subjective and objective quality assessment of video, IEEE Trans. Image Process. 19 (2010) 1427–1441.

[32] S. Tubaro, M. Tagliasacchi, A. Allam, P. Bestagini, S. Milani, Video codec identification, in: IEEE Int'l Conf. Acoust. Speech Signal Process., 2012.

[33] T. Brandao, M.P. Queluz, No-reference quality assessment of H.264/AVC encoded video, IEEE Trans. Circ. Syst. Video Technol. 20 (11) (Nov. 2010) 1437–1447.

[34] A. Eden, No-reference estimation of the coding PSNR for H.264-coded sequences, IEEE Trans. Consum. Electron. 53 (2) (2007) 667–674.

[35] P. Le Callet, C. Viard-Gaudin, D. Barba, A convolutional neural network approach for objective video quality assessment, IEEE Trans. Neural Netw. 17 (5) (2006) 1316–1327.

[36] K. Zhu, K. Hirakawa, V. Asari, D. Saupe, A no-reference video quality assessment based on laplacian pyramids, in: Proc. IEEE Int'l Conf. Image Process., Melbourne, 2013, pp. 49–53.

[37] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.

[38] M. Shahid, J. Panasiuk, G. Van Wallendael, M. Barkowsky, B. Lövström, Predicting full-reference video quality measures using HEVC bitstream-based no-reference features, in: IEEE Int'l Workshop on Quality of Multimedia Experience, Costa Navarino, Greece, May, 2015.

[39] M.A. Saad, A.C. Bovik, C. Charrier, Blind prediction of natural video quality, IEEE Trans. Image Process. 23 (3) (2014) 1352–1365.

[40] M. Narwaria, W. Lin, A. Liu, Low-complexity video quality assessment using temporal quality variations, IEEE Trans. Multimedia 14 (3) (2012) 525–535.

[41] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.) 67 (2) (2015) 301–320.

[42] J. Søgaard, S. Forchhammer, J. Korhonen, Video quality assessment and machine learning: performance and interpretability, in: IEEE Int'l Workshop on Quality of Multimedia Experience, Costa Navarino, Greece, May, 2015.

[43] F. Zhang, E. Steinbach, P. Zhang, MDVQM: a novel multidimensional no-reference video quality metric for video transcoding, J. Vis. Commun. Image Represent. 25 (3) (2014) 542–554.

[44] Y. Peng, E. Steinbach, A novel full-reference video quality metric and its application to wireless video transmission, in: Trans. IEEE Int'l Conf. on Image Processing, 2011, pp. 2517–2520.

[45] S.-H. Bae, J. Kim, M. Kim, S. Cho, J. Choi, Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services, IEEE Trans. Broadcast. 59 (2) (2013) 209–222.

[46] K. Sjöstrand, L. Clemmensen, R. Larsen, B. Ersbøll, Spasm: a matlab toolbox for sparse statistical modeling, J. Stat. Software (2012).

[47] J. Bi, K. Bennett, M. Embrechts, C. Breneman, M. Song, Dimensionality reduction via sparse support vector machines, J. Mach. Learn. Res. 3 (7–8) (2003) 1229–1243.

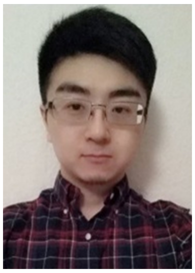
[48] L. Song, X. Tang, W. Zhang, X. Yang, P. Xia, The SJ TU 4K video sequence dataset, in: IEEE Int'l Workshop Quality Multimedia Experience, Klagenfurt, Austria, 2013.

[49] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.

[50] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.

[51] P.V. Vu, C.T. Vu, D.M. Chandler, A spatiotemporal most apparent-distortion model for video quality assessment, in: Proc. IEEE Int'l Conf. on Image Processing, 2011, pp. 2505–2508.

[52] L. Zhang, Y. Shen, H. Li, VSI: a visual saliency induced index for perceptual image quality assessment, IEEE Trans. Image Process. 23 (10) (2014) 4270–4281.



Xin Huang received the B.S. degree in Engineering from Xidian University, Xi'an, China, in 2004, and the M.S. and Ph.D. degrees from the Technical University of Denmark, Lyngby, Denmark, in 2006 and 2009, respectively. Currently, he is a Post-Doctoral Researcher with the Coding and Visual Communications Group, Technical University of Denmark. His current research interests include image and video coding, image and video processing, and error correcting codes.

Jacob Søgaard received the B.S. degree in Engineering, in 2010, and the M.S. degree in engineering, in 2012, from the Technical University of Denmark, Lyngby, where he is currently pursuing his Ph.D. degree with the Coding and Visual Communication group at the Department of Photonics. His research interests include image and video coding, image and video quality assessment, visual communication, and machine learning for Quality of Experience purposes.

Søren Forchhammer (M'04) received the M.S. degree in Engineering and the Ph.D. degree from the Technical University of Denmark, Lyngby, in 1984 and 1988, respectively. Currently, he is a Professor with DTU Fotonik, Technical University of Denmark, where he has been since 1988. He is Head of the Coding and Visual Communication Group at DTU Fotonik. His main interests include source coding, image and video coding, video quality, distributed video coding, processing for image displays, visual communications, communication theory and optical communications.