Erschienen in: IEEE Transactions on Circuits and Systems for Video Technology ; 25 (2015), 4. - S. 533-546
https://dx.doi.org/10.1109/TCSVT.2014.2363737

No-Reference Video Quality Assessment Based on Artifact Measurement and Statistical Analysis

Kongfeng Zhu, Chengqing Li, *Senior Member, IEEE*, Vijayan Asari, *Senior Member, IEEE*, and Dietmar Saupe

***Abstract*— A discrete cosine transform (DCT)-based no-reference video quality prediction model is proposed that measures artifacts and analyzes the statistics of compressed natural videos. The model has two stages: 1) distortion measurement and 2) nonlinear mapping. In the first stage, an unsigned ac band, three frequency bands, and two orientation bands are generated from the DCT coefficients of each decoded frame in a video sequence. Six efficient frame-level features are then extracted to quantify the distortion of natural scenes. In the second stage, each frame-level feature of all frames is transformed to a corresponding video-level feature via a temporal pooling, then a trained multilayer neural network takes all video-level features as inputs and outputs, a score as the predicted quality of the video sequence. The proposed method was tested on videos with various compression types, content, and resolution in four databases. We compared our model with a linear model, a support-vector-regression-based model, a state-of-the-art training-based model, and a four popular full-reference metrics. Detailed experimental results demonstrate that the results of the proposed method are highly correlated with the subjective assessments.**

***Index Terms*— Blocking artifact, discrete cosine transform (DCT), H.264/Advanced Video Coding (AVC), natural scene, no-reference (NR) measure, video quality assessment (VQA).**

I. Introduction

VIDEO services have been adopted widely in both mobile and fixed networks. To provide better service, the capability of digital cameras, smartphones, and tablet computers to acquire and display high-resolution images and videos continues to advance rapidly. However, the human appetite for electronic visual content is always high, and consumer demand is increasing rapidly [2]. Thus, content providers are interested in evaluating the performance of their services from the final users' perspective, i.e., their quality of experience (QoE). The QoE of the visual signal is of fundamental importance for numerous image and video processing applications, such as 3-D TV systems, surveillance systems, mobile video systems, and conferencing systems. Before multimedia data reach the final users, they pass through three main stages: 1) generation by a capture device; 2) compression with a codec; and 3) transmission via a communication channel. The most reliable way of assessing video quality is subjective evaluation, where a number of human users are asked to evaluate the perceived quality, summarized in mean opinion scores (MOSs). However, this approach is cumbersome, slow, and expensive for most applications.

In contrast, algorithms may provide an efficient and effective video quality assessment (VQA). VQA algorithms (or metrics) can be classified into three types according to how much reference information is used: 1) full-reference VQA (FR-VQA); 2) reduced-reference VQA (RR-VQA); and 3) no-reference VQA (NR-VQA) [2]. In some situations, such as evaluating the performance of digital camera and camcorder systems, the original uncorrupted images or videos are often unavailable because the imaging (sensing) and recording system are unknown and have to be treated as a black box, without providing access to the original video reference. For such applications, only NR-VQA is applicable. However, designing NR-VQA schemes for accurate prediction of visual quality is more difficult than for VQA with full or partial reference [3], [4].

Up to now, some work on NR-VQA has been reported to measure distortion of certain types. One straightforward approach is to run an NR-image quality assessment (NR-IQA) algorithm on video frames one by one. However, that approach does not perform well due to the lack of reference, the high complexity of the distortion in the videos, and the strong variation in the video content. Obviously, the low-level features in distortion measurements are significantly influenced by the content of the videos, and the relation between an objective distortion measurement and the (subjectively) perceived video quality remains unknown. Therefore, extracting content-independent features and exploring the unknown relation are two key problems to be solved to improve the performance of NR-VQA algorithms.

The previous two key problems are addressed by existing IQA and VQA algorithms in a two-stage framework that consists of distortion measurements followed by a nonlinear mapping [5], [6]. A distortion measurement quantifies the difference between the distorted data and the corresponding

Manuscript received March 5, 2014; revised June 10, 2014 and August 12, 2014; accepted September 30, 2014. Date of publication October 17, 2014; date of current version April 2, 2015. The work of C. Li was supported in part by the Alexander von Humboldt Foundation, Germany, and in part by the Hunan Provincial Natural Science Foundation of China (No. 2015JJ1013). This paper was recommended by Associate Editor P. Le Callet.

K. Zhu is with the Image and Video-Communications research group, University of Nantes, 44306 Nantes, France (e-mail: kongfeng.zhu@univ-nantes.fr).

D. Saupe is with the Department of Computer and Information Science, University of Konstanz, Konstanz 78457, Germany (e-mail: dietmar.saupe@uni-konstanz.de).

C. Li is with the College of Information Engineering, Xiangtan University, Xiangtan 411100, China (e-mail: chengqingg@gmail.com).

V. Asari is with the Department of Electrical and Computer Engineering, University of Dayton, Dayton, OH 45469 USA (e-mail: vasari1@udayton.edu).

reference. The nonlinear mapping is composed of one or more nonlinear functions that transform the collection of distortion measurements to a single score representing the overall perceived quality of the video.

There is a tradeoff between the complexity and the performance of these VQA algorithms. The VQA algorithms with good performance usually require a large number of features for the distortion measurement, and complicated training models for determining the nonlinear mapping. To achieve a balance between complexity and performance, this paper focuses on the distortion measurement, and proposes an NR-VQA algorithm with a small number of efficient and content-independent features and a simple strategy for the nonlinear mapping. The proposed algorithm is designed to assess the perceived quality of compressed videos, since most of the artifacts encountered in videos are a direct result of lossy compression.

Our video quality prediction is based on the analysis of discrete cosine transform (DCT) coefficients, frame-by-frame and without reference. This follows the previous two-stage framework. In the first stage, the distortion is quantified by combining the artifact measurements and a statistical analysis. Six feature maps are generated from the DCT coefficients of all 4×4 subblocks in the decoded frame. From the six bands, three features (sharpness, smoothness, and blockiness) are extracted to quantify the artifacts introduced by lossy compression, and three other features [kurtosis, mean Jensen–Shannon divergence (MJSD), and distribution noise] are calculated for the statistical analysis. In the second stage, temporal pooling transforms the frame-level features of all frames to six video-level features. Finally, from these six features, a trained multilayer neural network computes a single numerical value as the predicted video quality. Comprehensive experiments were conducted, showing the effectiveness of the proposed method.

The rest of this paper is organized as follows. In Section II, previous work on NR-VQA is reviewed. Then, we discuss the character of compressed video in the DCT-domain and motivate the choice of our features in Section III. Section IV details the proposed DCT-based NR-VQA model, including the generating bands, the extraction of the frame-level features, and the pooling of the features to predict the quality score. In Section V, we give the experimental results on four video databases and report on the correlation between the objective prediction with the subjective MOS. The conclusion is drawn and some avenues for further research are summarized in Section VI.

II. Previous Work

A large amount of work has been done to assess the quality of distorted images and videos by the two-stage framework, namely, distortion measurement and nonlinear mapping. In the following, we review previous work for both stages.

A. Distortion Measurement

1) Artifact Measurement: Assuming the video compression algorithm is known, for example, motion picture expert group (MPEG)-4 or H.264/Advanced Video Coding (AVC), distortion-specific NR-VQA algorithms can measure the specific artifacts that exist in the decoded video. Blockiness and blurriness (or lack of sharpness) are the most annoying artifacts and have received intensive attention. In the following section, we study blind IQA or VQA algorithms that measure one or more kinds of artifacts in the distorted image or video, and focus on the measurement of blockiness and blurriness.

Blurriness appears as a widening of edge width, thus a straightforward way is to measure the average edge width [7]. An indirect measurement is to analyze the statistics of the local edge gradients [8] and model the gradient image as a Markov chain [9]. Blurriness was also modeled as the loss of energy at high frequencies and measured from the local power of the high-frequency wavelet coefficients [10], the log-energy of the discrete wavelet transform subbands [11], and the image effective bandwidth [12]. In addition, both spectral and spatial properties of the image were explored to quantify the perceived sharpness [13].

Blockiness is an annoying impairment in a decoded image and video frames at low bit-rates. It originates from a block-based encoding. Thus, some NR blockiness measurement techniques model, the blocky image as a nonblocky image interfered by a pure blocky signal in the spatial domain, and then detect and evaluate the power of the pure blocky signal [14]–[16]. The detected blockiness was also weighted by models of the luminance and texture masking effects of the human visual system (HVS), and models of human perception, since the perception of blockiness is influenced by the amount of detail in the images and video [17]–[19]. However, these metrics are not efficient for H.264/AVC compressed videos because of the deblocking filter, which smoothes the sharp edges between macroblocks.

2) Statistical Analysis: Undistorted natural images are assumed to possess certain statistical properties that hold across different image contents [20], [21]. The natural scenes here refer to real environments, as opposed to laboratory stimuli, and may include human-made objects [21], thus any image or video obtained from a camera or camcorder is considered to be natural.

Based on the hypothesis that the presence of distortions in natural images alters the natural statistical properties of the images, researchers have attempted to develop general purpose NR-QA algorithms without prior knowledge of the specific types of distortion [22]. The natural image quality evaluator in [23] uses a simple and successful spacial-domain natural scene statistic (NSS) model to construct a quality aware collection of statistical features.

Another statistical approach is to estimate the peak signal-to-noise ratio (PSNR) of a compressed frame from the coded bitstream. The transform coefficients obtained from quantized coefficients have been variously conjectured to follow a Laplacian distribution, a Cauchy distribution, or a generalized Gaussian distribution [24]–[26]. For instance, in [25], the DCT coefficients were modeled using Cauchy and Laplace probability density functions, the maximum-likelihood estimation method then yielding an estimate of the coding error, which was weighted by the spatio-temporal

contrast sensitivity function of the HVS for the prediction of perceptual video quality.

B. Nonlinear Mapping

In previous work, a monotonic mapping function was usually applied to a quality measure to minimize the prediction error without changing the rank order. The simplest one is a linear function. There are also more sophisticated monotonic functions, such as a third-order polynomial function with monotonicity constraints [27], an S-shaped function [28], a four-parameter logistic function [29], and a five-parameter logistic function [30], [31]. Parameters of these functions are estimated by regression analysis between MOS and the corresponding quality measure in a database. The regression is useful and practical for algorithms based on one single distortion measure, e.g., PSNR, structural similarity (SSIM) [32], multi-scale structural similarity (MS-SSIM) [33], and visual information fidelity (VIF) [30].

Current VQA techniques tend to extract a large number of features for distortion measurement, so that the overall quality can be predicted more accurately by supervised learning methods based on subjective MOSs in a training set of images or video sequences. Two NR image quality measures were based on this two-stage framework [6], [34], extracting a set of low-level image features in image databases to learn a mapping from these features to subjective image quality scores. By formulating IQA as a pattern recognition problem, an FR-/RR-IQA metric was proposed based on 2-D mel-cepstrum for feature extraction and machine learning for feature pooling [5].

One popular choice to construct the nonlinear mapping from the distortion features to the perceived quality of the images or video has been neural networks [35]–[37]. For example, in [38], an NR-VQA method was presented based on nonlinear statistical modeling, where an ensemble of neural networks was used. Circular backpropagation neural networks were used in a methodology for the objective quality assessment of MPEG video streams to pool features extracted from bitstreams [39]. An RR-VQA algorithm was proposed based on a convolutional neural network, which allows a continuous time scoring of the video, and a time-delay neural network that integrates objective features along the temporal axis [40]. A general regression neural network was employed for the nonlinear mapping in NR-IQA. The features, including the mean value of the phase congruency image, the entropy of the phase congruency image, the entropy of the distorted image, and the gradient of the distorted image, were transformed to perceptual image quality via a neural network [41].

Support vector regression (SVR) is another popular option to determine the mapping from the extracted features to the subjective quality. It has been adopted for FR-IQA in [42] and for FR-VQA in [43]. A trained epsilon-SVR model was used in an NR-VQA algorithm to predict the video quality from the joint and marginal distributions of local wavelet coefficients [44].

Other machine learning methods have also been adopted for image and VQA lately. Partial least squares regression was used to calculate the weights of the features extracted from an H.264/AVC encoded bitstream [45]. The circular extreme learning machine (ELM), which is an augmented version of the basic ELM, handles the mapping of visual signals into quality scores in the RR-IQA metric in [46]. An NR bitstream-based objective video quality metric was constructed by genetic programming-based symbolic regression, which calculates reliable white-box models that allow one to determine the importance of the parameters [47].

III. Analysis of Compressed Natural Scenes

In this section, we analyze the appearance of distortion and the corresponding characters in compressed natural scenes, then introduce the decomposition of a natural image as a preprocessing step prior to the distortion measurement step.

A. Characteristics of Compressed Natural Scenes

The appearance of power laws in the power spectral densities of natural scenes [21] suggests that it is reasonable to assume that there exist statistical relations between the high-pass responses of natural images and their bandpass counterparts. Lossy video compression leads to distortion of the natural video, which usually manifests itself as a loss of texture and other image features in the high-frequency domain. Thus, lossy compression decreases the similarity between the different frequency bands of a natural image.

In the spatial domain, the loss of texture caused by compression appears as an increase of the smooth image area, in which pixel values are homogenous, and a decrease of the sharp image area, in which pixel values vary significantly from each other. In the DCT domain, lossy compression typically sets many ac coefficients to zero, thereby modifying the natural distributions, which were conjectured to be Gaussian, Laplacian, or Cauchy distributions [48]. In particular, the zero coefficients appear with a much higher probability.

An in-loop deblocking filtering technique has been adopted [49] to reduce blocking artifacts in H.264 compressed videos, but blockiness remains visible in the low-textured area. A new blockiness metric is needed to measure H.264 compressed videos due to the failure of existing blockiness measurements. We have found that blocking artifacts can be easily quantified based on the analysis of the horizontal and vertical DCT components.

In summary, the lossy compression of videos of natural scenes leads to an increase of the smooth image area, a decrease of the sharp image area, to the occurrence of blocking artifacts, a peaky ac coefficient distribution, and a dissimilarity between the bands of different frequencies.

B. Generation of Image Bands

To measure the distortion, a sliding window is moved over the decoded image pixel-by-pixel to generate six image bands $\mathbf{B}_1, \ldots, \mathbf{B}_6$. The window size is set to 4×4 for two reasons. First, it is the smallest size from which we can obtain three frequency bands while keeping the computational complexity as low as possible. Note that, the larger the size of the sliding window, the higher the computational complexity. Second,

C_1	C_2	C_3	C_4
C_5	C_6	C_7	C_8
C_9	C_{10}	C_{11}	C_{12}
C_{13}	C_{14}	C_{15}	C_{16}

Fig. 1. Coefficient names of a 4×4 DCT block.

Fig. 2. Decoded frame.

it is the right size to generate two orientation bands for blockiness measurement, when the transform block size is 4×4 in the H.264/AVC main profile [49]. Only the luminance is considered in our analysis, since the HVS is more sensitive to luminance than to chrominance (color) [49]. From the 16 DCT coefficients of the sliding window, we derive six features per pixel per frame, yielding a band per frame. Suppose, for example, that the frame size is $(M + 3) \times (N + 3)$. The generation of the six bands is described in the following four sections.

1) Generation of the DCT Coefficient Matrix: The 16 DCT coefficients of the sliding window are referred to as c_1 to c_{16} in raster order, as shown in Fig. 1. Letting the sliding window move over the whole frame, each coefficient yields a matrix $\mathbf{C}_i$, $i = 1, \ldots, 16$ of size $M \times N$. The matrix $\mathbf{C}_1$ contains dc coefficients, whereas the other 15 matrices contain the ac coefficients. As discussed in Section III-B, a lossy compression mainly affects the ac coefficients and further leads to the degradation of the quality of the compressed videos, thus only $\{\mathbf{C}_i\}_{i=2}^{16}$ are involved in the following calculation.

2) Generation of an Unsigned AC Band: Adding up the absolute values of the fifteen ac coefficients of a block, we get its unsigned ac band $\mathbf{B}_1$

$$\mathbf{B}_1(m,n) = \sum_{i=2}^{16} |\mathbf{C}_i(m,n)|$$

where $m = 1, \ldots, M$, and $n = 1, \ldots, N$. It will be used to measure smoothness, sharpness, and peakiness.

3) Normalization of the AC Coefficients: To normalize the DCT coefficients of a block, we divide each of its ac coefficients by its corresponding ac feature obtained in Section III-B2, and get 15 matrices $\{\tilde{\mathbf{C}}_i\}_{i=2}^{16}$, where

$$\tilde{\mathbf{C}}_i(m,n) = \frac{\mathbf{C}_i(m,n)}{\mathbf{B}_1(m,n)}$$

$m = 1, \ldots, M$, and $n = 1, \ldots, N$.

4) Generation of the Frequency and Orientation Bands: From these 15 matrices, we obtain three frequency bands $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4$ to quantify the dissimilarity and noise in their histograms, and two orientation bands $\mathbf{B}_5$ and $\mathbf{B}_6$ to quantify their blockiness. They are defined by

$$\mathbf{B}_2(m,n) = \sum_i \tilde{\mathbf{C}}_i(m,n), \quad i = 2, 5, 6$$

$$\mathbf{B}_3(m,n) = \sum_i \tilde{\mathbf{C}}_i(m,n), \quad i = 3, 7, 9, 10, 11$$

$$\mathbf{B}_4(m,n) = \sum_i \tilde{\mathbf{C}}_i(m,n), \quad i = 4, 8, 12, 13, \ldots, 16$$

$$\mathbf{B}_5(m,n) = \sum_i |\tilde{\mathbf{C}}_i(m,n)|, \quad i = 2, 3, 4$$

$$\mathbf{B}_6(m,n) = \sum_i |\tilde{\mathbf{C}}_i(m,n)|, \quad i = 5, 9, 13$$

where $m = 1, \ldots, M$ and $n = 1, \ldots, N$.

An example of a decoded frame is shown in Fig. 2 and the six bands of the corresponding frame are shown in Fig. 3. As demonstrated by Fig. 3, the six bands contain different information about the frame: 1) band $\mathbf{B}_1 \in [0, \infty]$ contains all the information of the image except the dc components of local regions; 2) bands $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4 \in [-1, 1]$ contain low-, medium-, and high-frequency components, respectively; and 3) bands $\mathbf{B}_5$ and $\mathbf{B}_6 \in [0, 1]$ contain vertical and horizontal components, respectively.

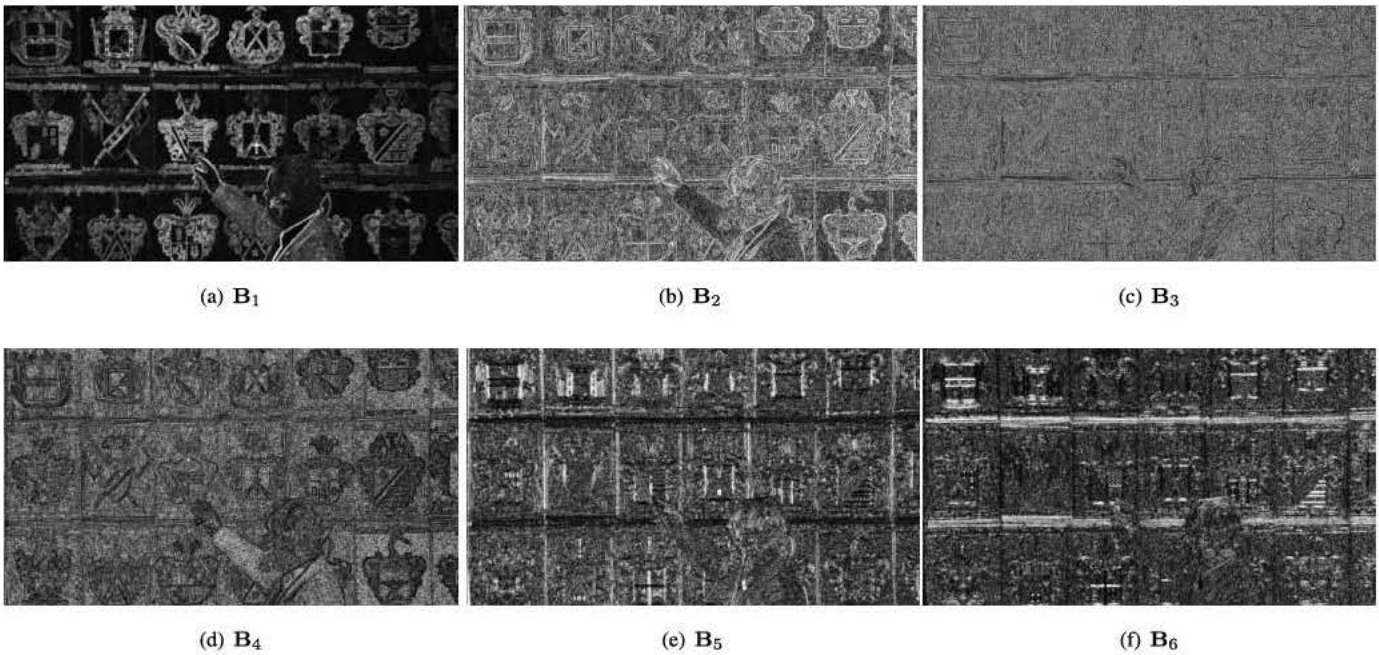
Table I lists the six generated bands. The kurtosis, smoothness, and sharpness will be computed on the unsigned ac band. Two statistical features, MJSD and histo-noise, will be extracted over the three frequency bands. We will measure the blocky artifacts on the two orientation bands. Note that the frequency bands $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4$ are signed rather than unsigned as in [1], because their histograms are bilaterally symmetric by keeping signs of their elements. The bilateral symmetry will increase the accuracy of difference measurements between their probability distributions, namely, the MJSD in Section IV.

IV. Proposed NR-VQA Algorithm

Based on the above analysis, we propose an NR-VQA algorithm in the two-stage framework: 1) extracting frame-level features and 2) video-level features from the six bands $\mathbf{B}_1$ to $\mathbf{B}_6$, followed by mapping the feature vectors to a quality prediction score by a neural network.

A. Frame-Level Feature Extraction

Based on the six bands, six frame-level features are extracted to quantify the distortion of compressed natural videos: 1) kurtosis; 2) smoothness; 3) sharpness; 4) MJSD; 5) histo-noise; and 6) blockiness. The features of kurtosis, smoothness, and sharpness quantify the distortion based on the statistical properties of band $\mathbf{B}_1$. Histo-noise and MJSD quantify the similarity between frequency bands based on the probability density functions of $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4$. To quantify



(a) $\mathbf{B}_1$ (b) $\mathbf{B}_2$ (c) $\mathbf{B}_3$

(d) $\mathbf{B}_4$ (e) $\mathbf{B}_5$ (f) $\mathbf{B}_6$

Fig. 3. Six bands $\mathbf{B}_1$ to $\mathbf{B}_6$ of the frame shown in Fig. 2. Sums of normalized magnitudes. (a) All 15 ac coefficients. (b) Low-frequency ac coefficients. (c) Medium-frequency ac coefficients. (d) High-frequency ac coefficients. (e) AC coefficients for vertical structures. (f) AC coefficients for horizontal structures.

TABLE I

LIST OF BANDS. THE FREQUENCY BANDS ARE SIGNED SUCH THAT THEIR HISTOGRAMS ARE SYMMETRIC

Band		Name	Range	Description	Index
Unsigned AC band		$\mathbf{B}_1$	$[0,\infty]$	$\sum_i \lvert\mathbf{C}_i\rvert$,	$i=2,\cdots,16$
Frequency	low	$\mathbf{B}_2$	$[-1,1]$	$\sum_i \tilde{\mathbf{C}}_i$,	$i=2,5,6$
	medium	$\mathbf{B}_3$	$[-1,1]$	$\sum_i \tilde{\mathbf{C}}_i$,	$i=3,7,9,10,11$
	high	$\mathbf{B}_4$	$[-1,1]$	$\sum_i \tilde{\mathbf{C}}_i$,	$i=4,8,12,\cdots,16$
Orientation	vertical	$\mathbf{B}_5$	$[0,1]$	$\sum_i \lvert\tilde{\mathbf{C}}_i\rvert$,	$i=2,3,4$
	horizontal	$\mathbf{B}_6$	$[0,1]$	$\sum_i \lvert\tilde{\mathbf{C}}_i\rvert$,	$i=5,9,13$

TABLE II

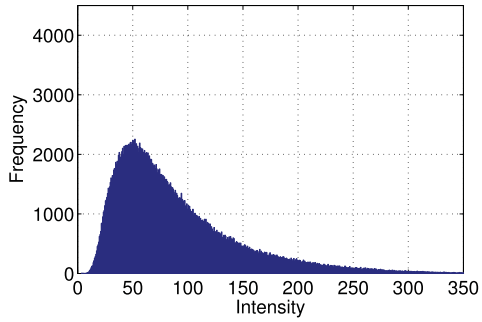
FEATURES OF FRAME t IN THE DCT-BASED MODEL. ALL THE FEATURES ARE BETWEEN 0 AND 1

Name	Feature	Range	Description
$f_1(t)$	peakiness	$(0,1]$	inverse of the kurtosis of unsigned AC band $\mathbf{B}_1$
$f_2(t)$	smoothness	$[0,1]$	relative sharp area of the current frame
$f_3(t)$	sharpness	$[0,1]$	relative edge area of the current frame
$f_4(t)$	MJSD	$[0,1]$	filtered distribution distance between $\mathbf{B}_2$, $\mathbf{B}_3$ and $\mathbf{B}_4$
$f_5(t)$	histo-noise	$[0,1]$	the average histogram noise of $\mathbf{B}_2$, $\mathbf{B}_3$ and $\mathbf{B}_4$
$f_6(t)$	blockiness	$(0,1]$	measurement of blocking artifacts

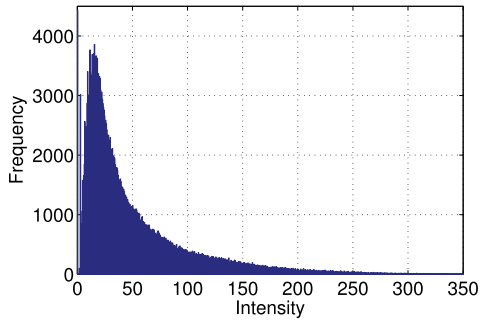
the blocking artifacts, the blockiness measurement is proposed based on $\mathbf{B}_5$ and $\mathbf{B}_6$.

Compared with our prior work presented in [1], we improve the feature extraction in many ways. First, the features of kurtosis ($\in [1,\infty)$) and blockiness ($\in [0,\infty)$) are remapped to $(0,1]$. The accuracy and robustness of the neural network will be improved when all features are in $(0,1]$ [50]. Second, the feature of MJSD between frequency bands is improved by generating $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4$ in a new manner, such that their filtered histograms are bilaterally symmetric. Together with the new feature of histo-noise, the modified MJSD quantifies the quality degradation better than that in [1]. These frame-level features are listed in Table II, and presented in more detail in the following.

1) Feature Extraction From the AC Band: The histograms of the summed unsigned ac band, $\mathbf{B}_1$, extracted from an original frame and the corresponding distorted frame are shown in Fig. 4. Compared with the original frame, the histogram for the distorted frame has a sharper main peak, an additional peak at near zero, and lower frequency at



(a)



(b)

Fig. 4. Histogram of the band $\mathbf{B}_1$ (summed unsigned ac coefficients) for intensities up to 300 and bin size equal to 0.5. (a) Original frame. (b) Distorted frame. This is an extreme example for videos with average quality. The compression rates of the two videos are 5 Mb/s and 200 kb/s, respectively.

high intensities. In probability theory and statistics, kurtosis measures the peakiness of the probability distribution of a real-valued random variable. Let $p_1(x)$ be the probability density functions of $\mathbf{B}_1$. We choose the inverse of the kurtosis as a feature: its value is within $(0, 1]$ and it is defined by

$$f_1(t) = \frac{\sigma_x^4}{E(x-\mu_x)^4} \in (0, 1] \tag{1}$$

where x is the intensity, μ_x is the mean of x, and σ_x is its standard deviation.

For each block, if the sum of the absolute ac coefficients is less than a given threshold T_L, it is considered to be a smooth block. The degree of smoothness is quantified by the relative area of the smooth region in the frame, which is defined as

$$f_2(t) = \frac{1}{\mathrm{MN}}\mathrm{card}(\{(m,n) \mid \mathbf{B}_1(m,n) < T_L\}) \in [0, 1] \tag{2}$$

where card (A) denotes the cardinality of a set A. The smoothness is expected to grow monotonically with respect to the compression ratio.

If the sum of the ac coefficients is greater than a given threshold T_H, the corresponding block is considered to be a sharp block. Sharpness is quantified as the relative area of the sharp region in the frame, defined by

$$f_3(t) = \frac{1}{\mathrm{MN}}\mathrm{card}(\{(m,n)|\mathbf{B}_1(m,n) > T_H\}) \in [0, 1]. \tag{3}$$

A compressed video with higher compression is expected to have a smaller sharp area.

2) Feature Extraction From the Frequency Bands: The bands $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4$ correspond to the low-, medium-, and high-frequency components. We assume that the frequency components of natural scenes are dependent and statistically smooth, and that a lossy compression reduces their dependence and statistical smoothness. Fig. 5(a)–(c) shows the histograms of the three bands $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4$ of an undistorted natural video frame and Fig. 5(d)–(f) shows the corresponding compressed video frame. The uncompressed frame exhibits a relatively smooth statistical distribution for each band, and there is some similarity in the distribution between bands, whereas the distribution for the compressed frame is noisier and shows less similarity between bands. Hence, two features, histo-noise and MJSD, are extracted to quantify the noise in the histograms and the dissimilarity between bands of different frequencies.

Write $\psi_i(x)$ for the noisy histogram of band $\mathbf{B}_i$ and $\bar{\psi}_i(x)$ for the filtered version of $\psi_i(x)$, where the median filter is adopted. The histogram noise of band $\mathbf{B}_i$ is defined by

$$\epsilon_i(x) = \frac{|\psi_i(x) - \bar{\psi}_i(x)|}{\sum_x \psi_i(x)}, \quad i = 2, 3, 4.$$

The histogram noise of the tth frame is defined as the mean of $\epsilon_i(x)$

$$f_4(t) = \frac{1}{3}\sum_x [\epsilon_2(x) + \epsilon_3(x) + \epsilon_4(x)] \in [0, 1]. \tag{4}$$

Define $p(x)$ and $q(x)$ as two probability mass functions. The Kullback–Leibler divergence (KLD) is a measure of the difference between two probability distributions and is given by

$$D_{\mathrm{KL}}(p||q) = \sum p(x) \log \frac{p(x)}{q(x)}.$$

The KLD is nonsymmetric. The symmetrized version of KLD is the Jensen–Shannon divergence (JSD), which is a symmetric measure of the distance between two probability distributions [51]. The JSD is defined as

$$D_{\mathrm{JS}}(p||q) = \frac{1}{2}(D_{\mathrm{KL}}(p||r) + D_{\mathrm{KL}}(q||r))$$

where $r(x) = (p(x) + q(x))/2$.

In Fig. 5, it can be observed that the similarity between two adjacent frequency bands of a natural video is decreased due to lossy compression. To measure the decrease of their similarity, the mean JSD of $\mathbf{B}_2, \mathbf{B}_3$, and $\mathbf{B}_4$ is defined as

$$f_5(t) = \frac{1}{2}\left(D_{\mathrm{JS}}(p_2||p_3) + D_{\mathrm{JS}}(p_3||p_4)\right) \in [0, 1] \tag{5}$$

where

$$p_i(x) = \frac{\bar{\psi}_i(x)}{\sum_x \bar{\psi}_i(x)}, \quad i = 2, 3, 4$$

i.e., $p_2(x)$, $p_3(x)$, and $p_4(x)$ are the smoothed probability density functions of $\mathbf{B}_2, \mathbf{B}_3$, and $\mathbf{B}_4$, respectively. In general, a high value of the MJSD means a low quality of the frame.

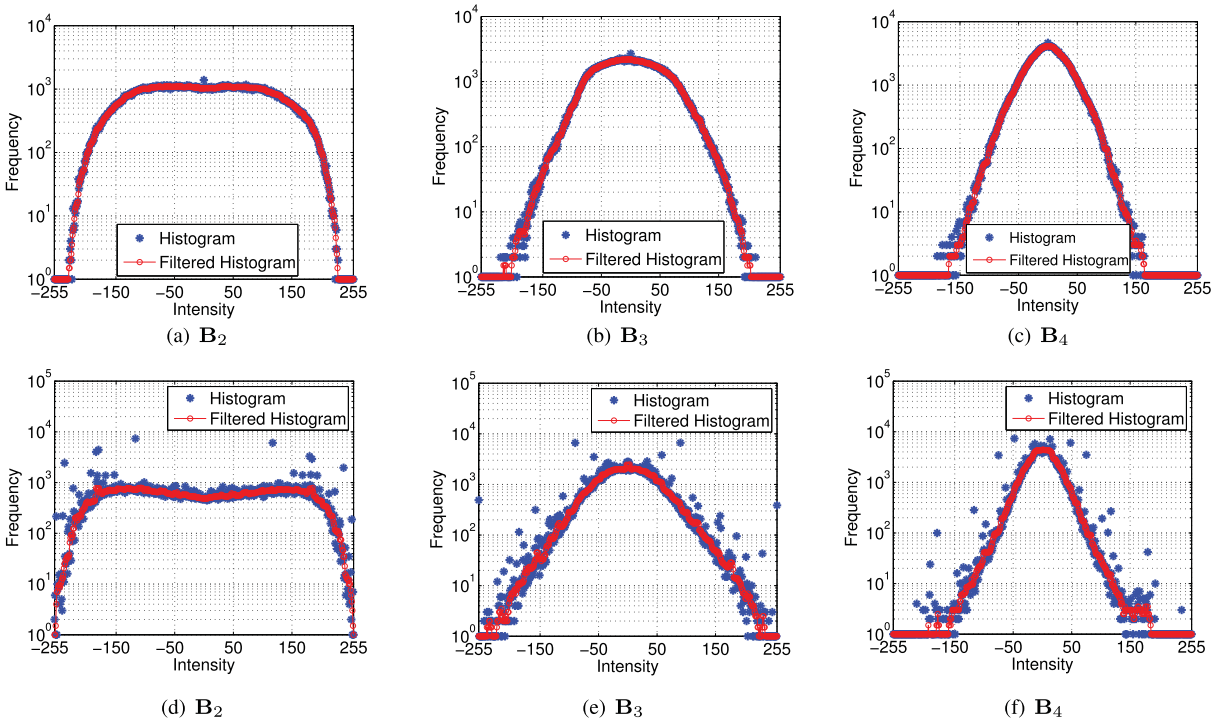


Fig. 5. Histograms of band $\mathbf{B}_2, \mathbf{B}_3$, and $\mathbf{B}_4$. (a)–(c) Undistorted video frame. (d)–(f) Compressed video frame. Histograms in (d)–(f) are very noisy in comparison to those in (a)–(c), whereas all the filtered histograms are roughly bilaterally symmetric.

3) Feature Extraction From the Orientation Bands: Due to the deficiencies of the existing blockiness metrics for H.264/AVC compressed videos, we propose a new metric to measure the blockiness in H.264/AVC compressed natural videos. In the new metric, the blocking artifacts are measured based on the two orientation bands of the decoded frame, rather than the decoded frame itself in previous methods. The horizontal and vertical blockiness are measured by applying the discrete Fourier transform (DFT), in the similar way as in [14]–[16], but on bands $\mathbf{B}_5$ and $\mathbf{B}_6$ rather than gradient images. The overall blockiness measurement is defined as the mean of the horizontal and vertical blockiness measurements. Empirically, the smaller the measurement value is, the worse the quality of the video will be. We describe the modified blockiness measurement as follows.

Assume the macroblock size of the codec is $S \times S$. We measure the horizontal blockiness by applying a sum operation along each row in band $\mathbf{B}_6$. This results in a 1-D array of length M, denoted by ϕ_H, where

$$\phi_H(m) = \sum_{n=0}^{N-1} \mathbf{B}_6(m,n), \quad m = 0, ..., M-1.$$

It is difficult to directly derive the blockiness power from ϕ_H. Fortunately, more clues can be obtained in the frequency domain [14]. We take the 1-D DFT of ϕ_H and consider the magnitude of the DFT coefficients, which can be expressed as

$$\Phi_H(l) = \left| \sum_{m=0}^{M-1} \phi_H(m) \exp\left(-\frac{j2\pi ml}{L}\right) \right|$$

where $l = 0, \ldots, L-1$ and L is the smallest power of 2 less than or equal to the upper limit M.

Due to the nature of the DFT, $\Phi_H(l)$ has peaks at $l = (L/S) \cdot s$, for $s = 1, 2, \ldots, S/2-1$. The values at those peaks are closely related to the horizontal blockiness of the image. The horizontal blockiness measurement is then computed as

$$P_H = \frac{1}{S/2-1} \sum_{s=1}^{S/2-1} \log_{10}\left(\Phi_H\left(\frac{L}{S} \cdot s\right) + 1\right) \in [0, \infty). \quad (6)$$

To scale the blockiness measurement to the interval $(0, 1]$, P_H is transformed to P_{LKH} by

$$P_{\mathrm{LKH}} = \frac{1}{1+P_H} \in (0, 1].$$

Fig. 6 shows the DFT coefficients of a reference frame and its corresponding distorted frame. A 512-point DFT was taken and the macroblock size was 16×16. No periodic peak is observed in the top subfigure, while periodic peaks appear at $l = 32, 64, 96, 128, 160, 192$, and 224 in the bottom subfigure. The values at these peaks are chosen for computing P_H in (6). Note that due to the symmetry of the DFT for real-valued signals, 14 peaks rather than seven are marked in each subfigure, but only the first seven peaks are used in the computation.

By applying a sum operation along each column in band $\mathbf{B}_5$, the vertical blockiness P_{LKV} is then measured accordingly. Finally, the overall blockiness is defined as

$$f_6(t) = \frac{1}{2}(P_{\mathrm{LKH}} + P_{\mathrm{LKV}}) \in (0, 1]. \quad (7)$$

B. Nonlinear Mapping

To predict the video quality from the frame-level features of all the frames of a video sequence, we nonlinearly map

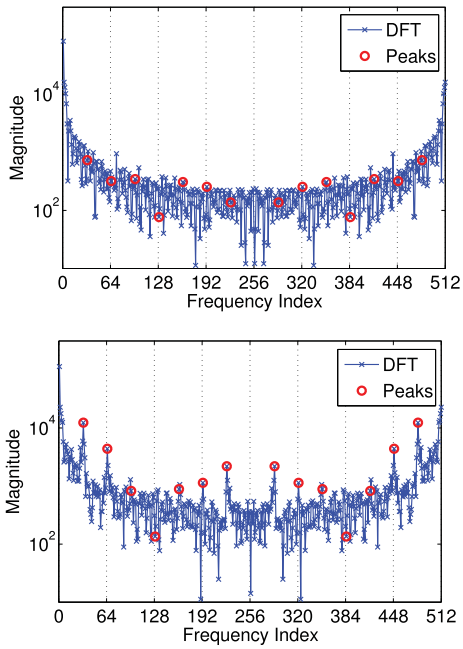


Fig. 6. DFT coefficients for the horizontal blockiness measurement.

the features to a score of the perceptual video quality. It is composed of a temporal pooling of the frame-level features and a multilayer neural network for combining the video-level features. For a video sequence, each frame-level feature yields a vector $(f_j(1), f_j(2), \ldots, f_j(t), \ldots, f_j(T_0))$, where T_0 is the total number of frames. The vector is then transformed to a video-level feature by Minkowski pooling

$$Q_j = \sqrt[4]{\frac{1}{T_0}\sum_{t=1}^{T_0} f_j^4(t)}$$

where $j = 1, 2, \ldots, 6$ [52].

The video-level features $Q_1, \ldots, Q_6$ are then treated as inputs to a neural network trained to predict the subjective video quality score. We choose the neural network rather than the popular SVR that performs better in general because the neural network can represents the predicted score as a parametric function of features with a fixed number of parameters. It is helpful to design a fixed parametric function for the nonlinear mapping in our future work. Eventually, an explicit nonlinear mapping will be developed to replace the black box based on machine learning. The model size of a kernel-based SVR, however, is not fixed in general, because the support vectors are selected from the training data, and the number varies according to the training data [53].

Fig. 7 gives the high-level organization of the proposed prediction model. It is composed of two stages. In the first stage, six bands are generated from the DCT coefficients. Second is a frame-level feature extraction stage, as described in Section IV-A and Table II. In the second stage, each extracted frame-level feature as a vector is first taken as an input to the temporal pooling. A single score results as the corresponding video-level feature along the time axis. An objective video quality score is then predicted by the trained neural network from the video-level features.

V. Performance Evaluation

Our VQA model is effective, as demonstrated by experiments on four video databases: 1) the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN) video database [54]; 2) the video quality experts group (VQEG) high-definition television (HDTV) Pool2 database [55]; 3) the LIVE mobile video database [56]; and 4) the LIVE video database [57]. These four databases contain many source videos with diverse content and resolution. Only compressed videos in the databases were used to evaluate its performance, since our method aims to objectively assess the quality of compressed videos.

A. Experimental Procedure

For all the experimental results in Section V-B, we set $T_L = 1$ in (2) and $T_H = 300$ in (3) for extracting frame-level features. A multilayer perceptron (MLP), which is a feed-forward artificial neural network model, was created with two layers, and 20 nodes were empirically set in the hidden layer for nonlinear mapping. The sigmoid transfer function was chosen for all hidden nodes and the output node, as all the input features are between 0 and 1. Since there are only six features and hundreds of videos in databases, we adopted the Levenberg–Marquardt backpropagation algorithm to update weight and bias values in the network during training [58]. This method is known to be fast and have stable convergence for small-size training problems.

Four statistical indices were used to evaluate the performance. They are the linear correlation coefficient (LCC) also known as Pearson's correlation coefficient, Spearman's rank-ordered correlation coefficient (SROCC), the root mean squared error (RMSE), and the mean absolute error (MAE) between the predicted quality scores and the MOS. A value close to 1 for the SROCC or LCC and a value close to 0 for the RMSE or the MAE indicates superior correlation with the subjective assessments. The four indices were defined in [31].

Depending on the size of the video database, we performed content-sensitive k-fold or leave p-fold-out cross-validation strategy to get a general performance of the proposed model. Let a denote the number of reference videos in a database, and let b be the number of distorted videos generated for each reference. In k-fold validation, the original videos were divided into k $(k \leq a)$ disjoint groups with equal size. Each group of original videos together with their distorted videos comprised one fold.

In one cross-validation process, a single fold is set for testing, and the remaining $(k-1)$ folds for training. The training and testing process is performed k times with a different fold of testing videos each time. This strategy is not robust when a and b are both very small. The leave p-fold-out strategy usually is preferred in such cases, namely, in one training and testing process, p folds are chosen for testing, and the remaining for training. Here, one-fold is composed of one original video and the corresponding distorted videos. The training and testing process is repeated $\binom{a}{p}$ times on $\binom{a}{p}$ train-test pairs.

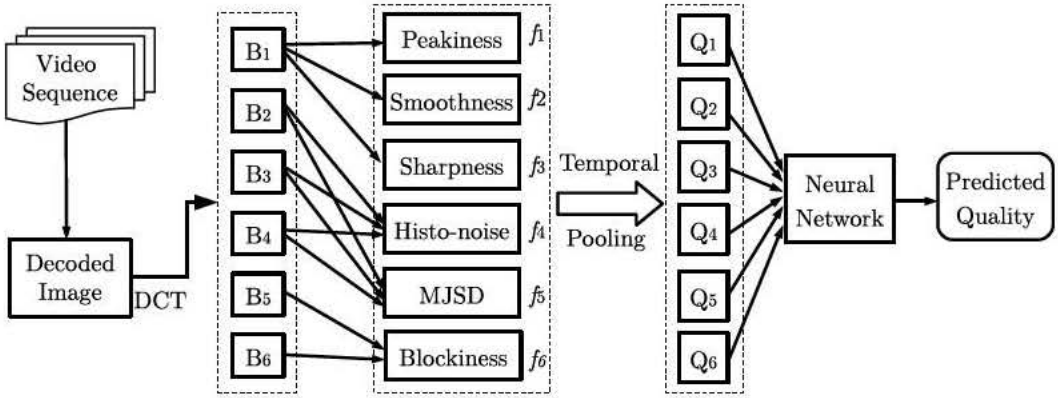


Fig. 7. Flowchart of the proposed model.

TABLE III

LIST OF VIDEO DATABASES FOR PERFORMANCE VALIDATION AND COMPARISON. THE a STANDS FOR THE NUMBER OF REFERENCES, b FOR THE NUMBER OF VIDEOS GENERATING FROM EACH REFERENCE WITH DIFFERENT QUALITY, AND ab FOR THE TOTAL NUMBER OF VIDEOS

Database	a	b	ab	Resolution	MOS range	Distortion	Cross-validation
IRCCyN video database	60	5	300	640×480	$[0, 5]$	H.264/SVC	10-fold
VQEG HDTV Pool2 database	9	8	72	1920×1080	$[0, 5]$	H.264, MPEG2	Leave-2-fold-out
LIVE mobile video database	10	4	40	1280×720	$[0, 5]$	H.264	Leave-2-fold-out
LIVE video database	10	8	80	768×432	$[0, 100]$	H.264, MPEG2	10-fold, leave-2-fold-out

We evaluated the performance of the proposed model in terms of the four statistical indices in testing data, and averaged their results from all testing processes to estimate the general performance of the proposed model in each video database. Their means and deviations are provided in Table IV. In addition to the validation of our VQA model on four databases, the nonlinear mapping based on the MLP was compared to the simple linear regression, and the same SVR model, as in [47], with a radial basis function kernel and the ε insensitive loss function. Schölkopf and Smola [53] and Haykin [58] for a detailed discussion about SVR. We also compared the performance with four popular FR-VQA metrics (PSNR, SSIM [32], MS-SSIM [33], and VIF [30]) on the IRCCyN video database, and a FR state-of-the-art VQA metric, presented in [43], on the LIVE video database.

B. Performance Evaluation and Comparison

For a fair comparison, we performed the cross validation with the same training and testing data for all models. Table III provides the basic information about the videos tested in our experiments and the strategies adopted for cross validation.

1) IRCCyN Video Database: This database contains 300 videos sequences of resolution 640×480 and the associated subjective results. There are 60 reference videos with different content and 5 versions for each one of them, including the original one and four distorted copies. Therefore, in the experiment $a = 60$ and $b = 5$. For each content, the reference and four distorted videos with random levels of degradation based on H.264/ scalable video coding (SVC) coding without transmission errors were subjectively evaluated. The absolute category rating was used as test methodology [54].

The 10-fold validation was performed, thus there were 30 videos for each test. The estimated performance was compared with linear regression and SVR in Table IV and Fig. 8(a). The means and standard deviations of the four indices indicated that all the three mapping methods can accurately and stably predict the video quality, and the MLP-based method gave the best performance. This is not only because the six extracted features are efficient for distortion measurement but also because of the large number of the videos and the single type of distortion in the database.

The performances of PSNR, SSIM [32], MS-SSIM [33], and VIF [30] are also given in Table IV and Fig. 8(a). Following the recommendation in [27], we chose the third-order polynomial function as their monotonic mapping function. Since only four parameters of the function were estimated on the training data, there is low risk of overfitting, as the lower standard deviation of the corresponding results suggest. In this case, the FR metrics are superior to NR-VQA metrics, as expected.

2) VQEG HDTV Pool2 Database: This is a full HD database. It is composed of 9 original sequences and 135 distorted videos by H.264 and MPEG2 coding with and without transmission error. The video resolution is 1920×1080 pixels at 59.94 fields/s in interlaced format [55].

We discarded videos with transmission errors and tested our model only on compressed videos, since the proposed model aims to predict quality of compressed videos. Hence, $a = 9$ and $b = 8$ in the experiment. In addition, we extracted features based on all fields rather than frames, because deinterlacing might introduce extra distortion to the video sequences. We performed leave two-fold-out validation with 16 videos compressed by H.264 and MPEG2 in each test. In comparison to the linear regression and the SVR in Table IV, the MLP still gave the best performance though the SVR was more precise (smaller standard deviations in Table IV). We also noticed that all the three mapping methods were not as stable as in the

TABLE IV

COMPARISON OF THE PERFORMANCE OF THE VQA ALGORITHMS. A VALUE CLOSE TO 1 FOR THE MEAN OF THE SROCC OR LCC, AND A VALUE CLOSE TO 0 FOR THE MEAN OF THE RMSE OR THE MAE INDICATES SUPERIOR CORRELATION WITH THE SUBJECTIVE ASSESSMENTS. A VALUE CLOSE TO 0 FOR THE STANDARD DEVIATION INDICATES A HIGH ROBUSTNESS OF THE CORRESPONDING MODEL

Database name (Validation strategy)	Model		Mean				Standard deviation			
			LCC	SROCC	RMSE	MAE	LCC	SROCC	RMSE	MAE
IRCCyN video database (10-fold)	Linear	(NR)	0.7112	0.7490	0.7892	0.6568	0.0951	0.0980	0.1012	0.0799
	SVR		0.7537	0.7825	0.7646	0.6338	0.0908	0.0871	0.0860	0.0695
	MLP		0.8302	0.8200	0.6179	0.4810	0.0808	0.0740	0.1359	0.0949
	PSNR	(FR)	0.8158	0.8638	0.6854	0.5468	0.0535	0.0466	0.0966	0.0731
	SSIM		0.8862	0.8884	0.5312	0.4252	0.0318	0.0338	0.0673	0.0561
	MS-SSIM		0.9345	0.9175	0.4038	0.3181	0.0278	0.0329	0.0766	0.0641
	VIF		0.9504	0.9247	0.3478	0.2782	0.0226	0.0328	0.0683	0.0541
VQEG HDTV Pool2 database (Leave-2-fold-out)	Linear		0.6898	0.7244	0.9866	0.8110	0.1493	0.1706	0.4315	0.3922
	SVR		0.7648	0.8037	0.7947	0.6682	0.1003	0.0994	0.0825	0.0667
	MLP		0.8459	0.8417	0.6018	0.4717	0.1471	0.1623	0.3106	0.2356
LIVE mobile video database (Leave-2-fold-out)	Linear		0.5005	0.5149	1.1491	1.0071	0.2454	0.2575	0.4489	0.4395
	SVR		0.4586	0.4371	1.1304	1.0086	0.3058	0.3154	0.1263	0.0962
	MLP		0.5856	0.6038	0.9541	0.8240	0.3141	0.3091	0.2999	0.2740
LIVE video database (Leave-2-fold-out)	Linear		0.3243	0.3652	10.5710	8.8455	0.2168	0.2410	2.6094	2.4339
	FR-VQA [43]		0.7720	0.7881	8.1426	6.8527	0.1003	0.1132	1.0406	0.9470
	Proposed NR-VQA		0.7855	0.8031	5.9897	4.5676	0.1704	0.1478	3.3688	2.8065
LIVE video database (10-fold)	Linear		0.5774	0.6731	8.1978	6.9122	0.2531	0.2570	1.7570	1.4078
	FR-VQA [43]		0.8348	0.8731	7.6561	6.4833	0.1053	0.0864	1.4758	1.4721
	Proposed NR-VQA		0.8354	0.8803	4.8778	3.8225	0.1727	0.1294	2.8704	2.3188

IRCCyN video database. This is not surprising since there are fewer references and two types of distortion in this database.

3) LIVE Mobile Video Database: It consists of 10 reference videos and 200 distorted videos, each of resolution 1280×720 at 30 frames/s, and with a duration of 15 s. A single-stimulus continuous quality evaluation study with hidden reference was conducted. The videos were displayed on the Motorola Atrix smartphone for a mobile study [56]. The quality ratings are in the range of $[0, 5]$.

For the same reason as in the VQEG HDTV Pool2 database, we validated our model on $ab = 40$, where $a = 10$ and $b = 4$, compressed videos. It is not suitable to perform $k-$fold cross validation with such small number of data. Therefore, the leave two-fold-out cross validation was carried out with 8 videos in each test, and the training and testing process was repeated 45 times. Unfortunately, all the mapping methods failed in the database. We attribute the failure to the small number of videos. The mapping was over trained on the training set, and therefore lead to a low and unreliable performance on testing set. The old metric in [1] was tested with the same cross-validation strategy and was indeed inferior to the proposed model (no data given in this paper).

4) LIVE Video Database: This database consists of 10 reference videos and 150 distorted videos, each with a resolution of 768×432 pixels and a length of 10 s. A total of 15 distorted sequences were generated from each reference sequence using four different distortion processes. Each distorted video was evaluated by 38 human observers. An MOS in the range of $[0, 100]$ is provided as the subjective quality assessment of each distorted video [57].

The leave 2-fold-out cross validation was performed on 80 ($a = 10$ and $b = 8$) videos compressed by MPEG-2 compression and H.264 compression. The training and testing process was repeated 45 times, each with 16 videos for testing. We compared the performance of our model with the training-based FR-VQA algorithm in [43]. The results of linear regression were also given to provide a baseline for readers.

The proposed model slightly outperformed the FR-VQA algorithm in [43] in terms of means of the LCC, SROCC, RMSE, and the MAE, but their standard deviations were higher in our model. Moreover, our model is distortion-specific and limited to compressed videos, whereas the FR-VQA in [43] is general purpose. However, we still believe the performance of our model is comparable with the FR-VQA algorithm since predicting video quality without reference is much harder.

We noticed that in our experiment the performance of FR-VQA [43] was worse than the original results reported in [43]. Assuming this is due to the different cross-validation strategy, we then conducted the 10-fold cross validation, as in [43]. As expected, we obtained similar results, as shown in Table IV.

The error bars of the LCC, SROCC, RMSE, and the MAE in all experiments were plotted in Fig. 8 to visually compare the performance of mapping methods and VQA models. In general, the more videos in a database, the better were the performance of our model. The linear mapping, with no surprise, performed worse than the SVR and the MLP. In addition, the performance of SVR-based nonlinear mapping was more precise (smaller standard deviations in Table IV)

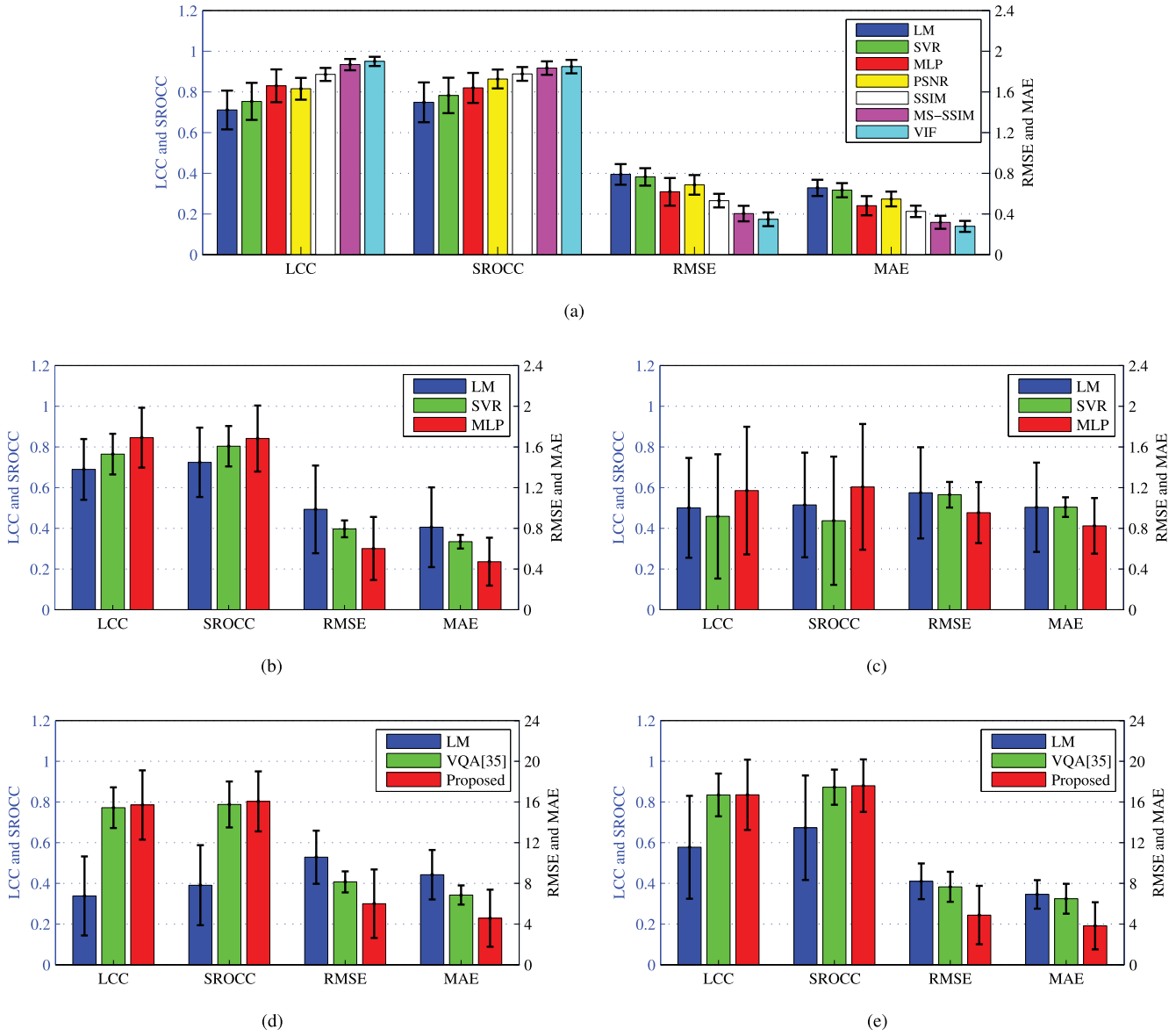


Fig. 8. Standard error bar for performance evaluation and comparison in (a) IRCCyN video database, (b) VQEG HDTV Pool2 database, (c) LIVE mobile video database, and (d) and (e) LIVE video database. The left y-axis corresponds to the LCC and SROCC, and the right y-axis to the RMES and MAE. The cross-validation strategies are 10-fold in (a) and (e), and leave two-fold out in (b)–(d).

throughout the cross validation than that of the MLP-based mapping, though the former was not as accurate as the latter. Therefore, it is hard to conclude, which one is better than the other.

5) Summary: Instead of adopting the leave-one-out cross-validation approach for only one limited size video database, as in [1], we have conducted a more comprehensive validation. Experiments used four video databases with diverse content and various types of video compression, and performed content-sensitive cross-validation strategies to estimate the general performance on unknown data. The proposed model was also compared with training-based methods, such as the linear regression and the SVR, and benchmarked against four popular FR-VQA metrics and a state-of-the-art training-based FR-VQA algorithm. In addition to improving the algorithm performance from [1], the statistical results presented in Table IV and Fig. 8 provide a more robust validation of the model.

VI. Conclusion

The VQA by subjective user studies is time consuming and expensive and may be replaced by a suitably designed objective NR-VQA. To assess the quality of H.264 coded videos, an NR-VQA model was presented, based on analyzing the local DCT coefficients of compressed video frames. Based on the properties of natural scenes and the different types of distortion in compressed videos, the proposed model combines the existing artifact- and NSS-based approaches. It is comparable with one state-of-the-art FR-VQA method according to the evaluation results for the LIVE video database.

The proposed model can quantify the distortion of a video sequence by extracting a few but efficient features for

distortion measurement and adopting a simple neural network for the quality prediction. In the model, a DCT was taken within a local window, which moves pixel-by-pixel over the entire frame to generate the so-called DCT map. For each frame, six bands are extracted from the DCT map. Six frame-level features, including three artifact metrics and three statistical metrics, are extracted from these bands. The frame-level features are transformed to video-level features through temporal pooling. Finally, a trained multilayer neural network gives the predicted video quality according to the six video-level features.

The performance evaluation was conducted on the IRCCyN video database, the VQEG HDTV Pool2 database, the LIVE mobile video database, and the LIVE video database. The results for videos compressed by H.264/SVC, H.264, and MPEG2 show a strong correlation between the predicted quality and the subjectively assessed quality. It is also clear that the proposed model is comparable to one FR-VQA algorithm in terms of Pearson's correlation coefficient, SROCC, the RMSE, and the MAE.

However, the proposed metric is distortion specific and data driven. Thus, the disadvantage of data-driven approaches is also applied to our model. For example, it is highly prone to overfitting of their parameters on small training sets, and therefore results in a low performance on unknown data. In addition, tests of their general performance on unknown data are not robust across content. Moreover, the uncertainty of subjective results [27] and the uncertainty of the predicted quality scores call for better statistical tools to evaluate and compare the performance of machine-learning-based methods.

The proposed model is designed for compression-distorted videos and aims at evaluating the performance of imaging systems based on the H.264/SVC, H.264, or MPEG2 compression standard, such as mobile phone cameras, HD camcorders, and video surveillance systems. Thus, its application is limited to compressed videos. The luma component was used for measuring the distortion frame by frame during the video analysis. Hence, the distortion in the temporal domain and the chroma components cannot be obtained using the proposed method. To include these effects, further study of the properties of natural scenes and the influence of various compressions on these properties is required. In addition, it may be possible to identify the most efficient features among the extracted ones to simplify the nonlinear mapping model and to decrease the complexity of the model, whereas preserving or improving its performance. An explicit nonlinear mapping, for instance, a parametric function of features, is also in demand to circumvent the intractable problem of overfitting in data-driven methods.

Acknowledgment

The authors would like to thank the anonymous reviewers for their detailed and helpful comments and suggestions.

References

[1] K. Zhu, V. Asari, and D. Saupe, "No-reference quality assessment of H.264/AVC encoded video based on natural scene features," *Proc. SPIE*, vol. 8755, no. 4, p. 875505, May 2013.

[2] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

[3] K. Zhu, S. Li, and D. Saupe, "An objective method of measuring texture preservation for camcorder performance evaluation," *Proc. SPIE*, vol. 8293, no. 6, p. 829304, Jan. 2012.

[4] K. Zhu and D. Saupe, "Performance evaluation of HD camcorders: Measuring texture distortions using Gabor filters and spatio-velocity CSF," *Proc. SPIE*, vol. 8653, no. 9, p. 86530A, Feb. 2013.

[5] M. Narwaria, W. Lin, and A. E. Çetin, "Scalable image quality assessment with 2D mel-cepstrum and machine learning approach," *Pattern Recognit.*, vol. 45, no. 1, pp. 299–313, Jan. 2012.

[6] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[7] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. Int. Conf. Image Process.*, vol. 3. 2002, pp. 57–60.

[8] C. Feichtenhofer, H. Fassold, and P. Schallauer, "A perceptual image sharpness metric based on local edge gradient analysis," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 379–382, Apr. 2013.

[9] C. Chen, W. Chen, and J. A. Bloom, "A universal reference-free blurriness measure," *Proc. SPIE*, vol. 7867, Jan. 2011, Art. ID 78670B.

[10] N. N. Ponomarenko, V. V. Lukin, O. I. Eremeev, K. O. Egiazarian, and J. T. Astola, "Sharpness metric for no-reference image visual quality assessment," *Proc. SPIE*, vol. 8295, Jan. 2012, Art. ID 829519.

[11] P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Process. Lett.*, vol. 19, no. 7, pp. 423–426, Jul. 2012.

[12] B. Fishbain, L. Yaroslavsky, I. Ideses, and F. Roffet-Crete, "No-reference method for image effective bandwidth estimation," *Proc. SPIE*, vol. 6808, Jan. 2008, Art. ID 68080X.

[13] C. T. Vu, T. D. Phan, and D. M. Chandler, "S_3: A spectral and spatial measure of local perceived sharpness in natural images," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 934–945, Mar. 2012.

[14] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proc. Int. Conf. Image Process.*, vol. 3. Vancouver, BC, Canada, 2000, pp. 981–984.

[15] H. Liu and I. Heynderickx, "A no-reference perceptual blockiness metric," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Mar./Apr. 2008, pp. 865–868.

[16] C. Chen and J. A. Bloom, "A blind reference-free blockiness measure," in *Proc. 11th Pacific Rim Conf. Adv. Multimedia Inf. Process.*, 2010, pp. 112–123.

[17] G. Zhai, W. Zhang, X. Yang, W. Lin, and Y. Xu, "No-reference noticeable blockiness estimation in images," *Signal Process., Image Commun.*, vol. 23, no. 6, pp. 417–432, Jul. 2008.

[18] H. Liu and I. Heynderickx, "A perceptually relevant no-reference blockiness metric based on local image characteristics," *EURASIP J. Adv. Signal Process.l*, vol. 2009, Jan. 2009, Art. ID 263540.

[19] L. Abate, G. Ramponi, and J. Stessen, "Detection and measurement of the blocking artifact in decoded video frames," *Signal Image Video Process.*, vol. 7, no. 3, pp. 453–466, May 2013.

[20] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, pp. 1193–1216, May 2001.

[21] W. S. Geisler, "Visual perception and the statistical properties of natural scenes," *Annu. Rev. Psychol.*, vol. 59, pp. 167–192, Aug. 2007.

[22] R. Soundararajan and A. C. Bovik, "Survey of information theory in visual quality assessment," *Signal, Image Video Process.*, vol. 3, no. 3, pp. 391–401, May 2013.

[23] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[24] H. Li and S. Forchhammer, "MPEG2 video parameter and no reference PSNR estimation," in *Proc. Picture Coding Symp.*, Chicago, IL, USA, May 2009, pp. 1–4.

[25] T. Brandão and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.

[26] J.-W. Ryu, S.-O. Lee, D.-G. Sim, and J.-K. Han, "No-reference peak signal to noise ratio estimation based on generalized Gaussian modeling of transform coefficient distributions," *Opt. Eng.*, vol. 51, no. 2, Feb. 2012, Art. ID 027401.

[27] *Recommendation ITU-T P.1401: Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, document International Telecommunication Union, Jul. 2012.

[28] *Recommendation ITU-T J.341: Objective Perceptual Multimedia Video Quality Measurement of HDTV for Digital Cable Television in the Presence of a Full Reference*, document International Telecommunication Union, Jan. 2011.

[29] A. M. Rohaly *et al.*, "Video quality experts group: Current results and future directions," *Proc. SPIE*, vol. 4067, pp. 742–753, May 2000.

[30] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[31] L. Cui and A. R. Allen, "An image quality metric based on corner, edge and symmetry maps," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–10.

[32] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[33] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.

[34] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 305–312.

[35] X. Jiang, F. Meng, J. Xu, and W. Zhou, "No-reference perceptual video quality measurement for high definition videos based on an artificial neural network," in *Proc. IEEE Int. Conf. Comput. Elect. Eng.*, Phuket, Thailand, Dec. 2008, pp. 424–427.

[36] D. Culibrk, D. Kukolj, P. Vasiljević, M. Pokrić, and V. Zlokolica, "Feature selection for neural-network based no-reference video quality assessment," in *Proc. 19th Int. Conf. Artif. Neural Netw.*, Limmassol, Cyprus, Sep. 2009, pp. 633–642.

[37] J. Choe, J. Lee, and C. Lee, "No-reference video quality measurement using neural networks," in *Proc. 16th Int. Conf. Digital Signal Process.*, Santorini, Greece, Jul. 2009, pp. 1–4.

[38] D. D. Kukolj, M. Pokrić, V. M. Zlokolica, J. Filipović, and M. Temerinac, "No-reference video quality assessment design framework based on modular neural networks," in *Proc. Int. Conf. Artif. Neural Netw.*, vol. 6352. Thessaloniki, Greece, Sep. 2010, pp. 569–574.

[39] P. Gastaldo, S. Rovetta, and R. Zunino, "Objective quality assessment of MPEG-2 video streams by using CBP neural networks," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 939–947, Jul. 2002.

[40] P. Le Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1316–1327, Sep. 2006.

[41] C. Li, A. C. Bovik, and X. Wu, "Blind image quality assessment using a general regression neural network," *IEEE Trans. Neural Netw.*, vol. 22, no. 5, pp. 793–799, May 2011.

[42] M. Narwaria and W. Lin, "Objective image quality assessment based on support vector regression," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 515–519, Mar. 2010.

[43] M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 525–535, Jun. 2012.

[44] M. Dimitrievski and Z. Ivanovski, "Fusion of local degradation features for no-reference video quality assessment," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Santander, Spain, Sep. 2012, pp. 1–6.

[45] Z. Shi, P. Chen, C. Feng, L. Huang, and W. Xu, "Research on quality assessment metric based on H.264/AVC bitstream," in *Proc. Int. Conf. Anti-Counterfeiting, Security Identificat. (ASID)*, Taipei, Taiwan, Aug. 2012, pp. 1–5.

[46] S. Decherchi, P. Gastaldo, R. Zunino, E. Cambria, and J. Redi, "Circular-ELM for the reduced-reference assessment of perceived image quality," *Neurocomputing*, vol. 102, pp. 78–89, Feb. 2013.

[47] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, "Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1322–1333, Aug. 2013.

[48] Y. Altunbasak and N. Kamaci, "An analysis of the DCT coefficient distribution with the H.264 video coder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 3. Montreal, QC, Canada, May 2004, pp. 177–180.

[49] *Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification*, document ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC, Geneva, Switzerland, May 2003.

[50] R. Rojas, *Neural Networks—A Systematic Introduction*. Berlin, Germany: Springer-Verlag, 1996.

[51] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[52] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, 1st ed. San Mateo, CA, USA: Morgan Kaufmann, 2006.

[53] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[54] Y. Pitrey, M. Barkowsky, R. Pépion, P. Le Callet, and H. Hlavacs, "Influence of the source content and encoding configuration on the perceived quality for scalable video coding," *Proc. SPIE*, vol. 8291, p. 82911K, Feb. 2012. [Online]. Available: http://130.66.64.103/spip.php?article771

[55] M. Barkowsky, M. Pinson, R. Pépion, and P. Le Callet, "Influence of the source content and encoding configuration on the perceived quality for scalable video coding," in *Proc. 5th Int. Workshop Video Process. Quality Metrics*, pp. 1–6, Jan. 2010. [Online]. Available: http://130.66.64.103/spip.php?article775

[56] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[57] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[58] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1999.

Kongfeng Zhu was born in Shandong, China. She received the M.Sc. degree in communication and information systems from Shandong University, Jinan, China, in 2006 and the Ph.D. degree in multimedia signal processing from University of Konstanz, Konstanz, Germany, in 2014. Her Ph.D. thesis focuses on no-reference video quality assessment and applications.

She was a Lecturer with Anhui University, Hefei, China, from 2006 to 2010. From 2010 to 2014 she was with the University of Konstanz. She is currently a Post-Doctoral Fellow with the University of Nantes, Nantes, France. Her research interests include visual quality assessment, image processing, motion analysis, feature selection, and machine learning.

Chengqing Li (M'07–SM'13) was born in Xiangxiang, China. He received the bachelor's degree in mathematics from Xiangtan University, Xiangtan, China; the M.Sc. degree in applied mathematics from Zhejiang University, Hangzhou, China, in 2005; and the Ph.D. degree in electronic engineering from City University of Hong Kong, Hong Kong, in 2008.

He was a Post-Doctoral Fellow with Hong Kong Polytechnic University, Hong Kong. From 2013 to 2014 he was with University of Konstanz, Konstanz, Germany, under support of Alexander von Humboldt Foundation. Since 2010 he has been with the College of Information Engineering, Xiangtan University, as an Associate Professor. He focuses on security analysis of image encryption schemes, and has authored 30 papers on the topic for the past nine years.

Vijayan Asari (M'96–SM'01) received the bachelor's degree in electronics and communication engineering from University of Kerala, Thiruvananthapuram, India, in 1978 and the M.Tech. and Ph.D. degrees in electrical engineering from IIT Madras, Chennai, India, in 1984 and 1994, respectively.

He was a Professor of Electrical and Computer Engineering with Old Dominion University (ODU), Norfolk, VA, USA; a Research Fellow with the National University of Singapore, Singapore, and with Nanyang Technological University, Singapore; and an Assistant Professor with University of Kerala. He is currently a Professor of Electrical and Computer Engineering and the Ohio Research Scholars Endowed Chair in Wide Area Surveillance with University of Dayton (UD), Dayton, OH, USA. He is also the Director of UD Vision Laboratory with the Center of Excellence for Computer Vision and Wide Area Surveillance Research, UD. He has authored over 440 research papers, including 76 peer-reviewed journal papers in image processing, computer vision, machine learning, pattern recognition, and high-performance digital architectures, and holds three patents. He has supervised 17 Doctoral Dissertations and 32 master's theses since joining ODU in 2000.

Dr. Asari is a Senior Member of the International Society for Optics and Photonics. He received several teaching, research, and advising awards.

Dietmar Saupe received the Ph.D. and Habilitation degrees in mathematics from University of Bremen, Bremen, Germany.

He held professorships with University of California at Santa Cruz, Santa Cruz, CA, USA, from 1985 to 1987; University of Freiburg, Freiburg im Breisgau, Germany, from 1993 to 1998; and Leipzig University, Leipzig, Germany, from 1998 to 2002. He is currently a Professor of Computer Science with University of Konstanz, Konstanz, Germany, where he leads the Multimedia Signal Processing Group. His research interests include multimedia signal processing, sport informatics, scientific visualization, and image processing.