No-reference quality assessment of H.264/AVC encoded video based on natural scene features

Kongfeng Zhu[a,b], Vijayan Asari[b], Dietmar Saupe[a]

[a]Department of Computer and Information Science, University of Konstanz, Germany
[b]Department of Electrical and Computer Engineering, University of Dayton, OH, USA

ABSTRACT

H.264/AVC coded video quality is crucial for evaluating the performance of consumer-level video camcorders and mobile phones. In this paper, a DCT-based video quality prediction model (DVQPM) is proposed to blindly predict the quality of compressed natural videos. The model is frame-based and composed of three steps. First, each decoded frame of the video sequence is decomposed into six feature maps based on the DCT coefficients. Then five efficient frame-level features (kurtosis, smoothness, sharpness, mean Jensen Shannon divergence, and blockiness) are extracted to quantify the distortion of natural scenes due to lossy compression. In the last step, each frame-level feature is averaged across all frames (temporal pooling); a trained multilayer neural network takes the five features as inputs and outputs a single number as the predicted video quality. The DVQPM model was trained and tested on the H.264 videos in the LIVE Video Database. Results show that the objective assessment of the proposed model has a strong correlation with the subjective assessment.

Keywords: Video quality assessment, no-reference, H.264/AVC, natural scenes, DCT, blockiness

1. INTRODUCTION

The video compression standard H.264, which is also known as MPEG-4 Part 10/AVC for Advanced Video Coding, has become the video standard of choice. It is jointly defined by standardization organizations in the telecommunications and IT industries, and has been more widely adopted than previous video compression standards, such as Motion JPEG and MPEG-4 Part 2.

H.264 has already been widely accepted in mobile phones, digital video players, HD camcorders, video surveillance systems, online video storage, and so on. Though H.264 provides savings in network bandwidth and storage costs, the increasing demands for image/video capture, transmission, and storage are still not satisfied by the limited bandwidth. The increasing demands also lead to the occurrence of information loss and extraneous artifacts.

How the distortions affect the quality of viewing experience has become the interest of researchers in visual quality assessment. Naturally, the subjective assessment is the golden standard, however, it is time-consuming, cumbersome and impractical. Hence one seeks to develop algorithms that produce objective quality estimates of these distorted visual stimuli with high correlation with subjective assessment. This paper aims to present a novel model for objective quality assessment of H.264 coded videos.

The rest of this paper is organized as follows. In Section 2, we review previous work in image/video quality assessment. We analyze the DCT-domain features of compressed videos and the motivation behind the choice of the features in Section 3. In Section 4, we show how the features are extracted in detail. The framework of the proposed prediction model is given in Section 5. In Section 6, we present the result on the LIVE video database and report how the objective prediction correlates with subjective Mean Opinion Scores (MOS). Conclusions and our future work are presented in Section 7.

Contact information of the authors: Kongfeng Zhu: Kongfeng.Zhu@uni-konstanz.de; Vijayan Asari: vasari1@udayton.edu; Dietmar Saupe: Dietmar.Saupe@uni-konstanz.de.

2. PREVIOUS WORK

Objective quality assessment can be divided into three categories depending on the amount of information provided to the algorithm. Full-reference (FR) algorithms are provided with the original undistorted visual stimulus along with the distorted stimulus whose quality is to be assessed. Reduced reference (RR) approaches are those in which the algorithm is provided with the distorted stimulus and some additional information about the original stimulus, either by using an auxiliary channel or by incorporating some information in the distorted stimulus (such as a watermark). No-reference (NR) approaches to quality assessment are those in which the algorithm is provided only with the distorted stimulus.

A large number of FR video quality assessment (VQA) algorithms have been proposed by several researchers. Some of them have become influential in the area and been used in practice. Multi-scale structural similarity (MS-SSIM)[1,2] is an extension of the widely accepted SSIM paradigm, and has been adopted for VQA by applying it frame-by-frame on the luminance component of the video.[3] Video Quality Metric (VQM) is a VQA algorithm adopted by the American National Standards Institute (ANSI) as a national standard, and as International Telecommunications Union Recommendations ITU-T J.144 and ITU-R BT.1683.[4] V-VIF is an extension of the Visual Information Fidelity (VIF) criterion for still images proposed to video using temporal derivatives.[5] The MOtion-based Video Integrity Evaluation (MOVIE) index is another excellent VQA index that was recently developed.[6]

Without a priori knowledge about the pristine image or video, NR image/video quality assessment (IQA/VQA) is the most useful but also the most difficult one to accurately predict visual quality. There are two main methods to assess image quality. The artifact-based method quantifies a specific distortion (for example blocking, blurring, and ringing), and scores an image accordingly. It is application specific and the number of distortions introduced to images in a wide range of applications is large, thus it is difficult for an algorithm to comprehensively quantify every type of distortion possible. The NSS-based method intends to capture the Natural Scenes Statistics (NSS), and take them as reference to quantify how much the scenes have been distorted in the image/video. It is general-purpose, and extracting content-independent statistical features of natural scenes is of great importance.

Blockiness measurement, as a part of artifact-based method, has attracted attention, because blocking artifacts are introduced by compression techniques based on the block discrete cosine transform (DCT) and have strong influence on the overall perceptual quality. Most of the NR blockiness measurement techniques model the blocky image as a non-blocky image interfered with a pure blocky signal in the spatial domain, and then the power of the pure blocky signal is detected and evaluated differently.[7–10] They can accurately assess the blockiness in image/video compressed by traditional standards, such as JPEG, H.261, H.263, MPEG-1, MPEG-2 and MPEG-4. However, all of them failed for H.264 compressed videos because the blocky image model in H.264 is invalid when in-loop deblocking filtering is employed to eliminate blocking artifacts.

Aiming to develop general-purpose NR algorithms, a great effort has been made based on the statistics of natural images.[11–13] Natural undistorted images are considered to possess certain statistical properties that hold across different image contents. The natural scene statistic (NSS) models are based on the hypothesis that the presence of distortions in natural images alters the natural statistical properties of images. The natural scenes here refer to real environments, as opposed to laboratory stimuli, and may include human-made objects,[12,14] thus any image or video that can be obtained from a camera or camcorder is considered to be natural.

The Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index is an NR IQA algorithm that popularized image quality assessment based on NSS.[15] It is capable of assessing the quality of a distorted image across multiple distortion categories (in contrast to most NR IQA algorithms that are distortion-specific). In another recent work, an extended peak signal-to-noise ratio (PSNR) metric was proposed for digital video subject to H.264/AVC encoding.[16] Assuming that DCT coefficients of natural scenes are corrupted by quantization noise, the coding error was estimated first, then a spatio-temporal contrast sensitivity function was applied to the DCT domain to perceptually weight the estimated coding error. We name this no-reference VQA metric as XT-PSNR, where XT stand for "extended".

Inspired by previously proposed models for NR IQA and VQA, in the paper we combine both artifact-based and NNS-based methods to objectively predict the quality of H.264/AVC coded videos. A DCT-based video

(a) Original frame (b) Distorted frame

Figure 1. Blockiness in H.264 compressed videos

quality prediction model (DVQPM) is presented to blindly predict the quality of encoded video sequences frame by frame. Six feature maps are generated from the DCT coefficients of all 4×4 subblocks in the decoded frame. We extract three features (sharpness, smoothness, and blockiness) to assess artifacts introduced by compression and two other statistical features (kurtosis and mean Jensen Shannon divergence). A temporal pooling strategy is adopted to pool frame-level features of all frames to five video-level features; a trained multilayer neural network, taking the five video-level features as inputs, outputs a single value as the predicted video quality.

In the proposed method, much fewer features were extracted than for that in Ref. 15 and the undistorted reference videos used in XT-PSNR [16] for training are not required. Moreover, the training procedures in Refs. 15, 16 are complex. We simplify the training procedure by adopting a two-layer neural network. The performance of DVQPM is evaluated on the LIVE video databases. Results show that the proposed model outperforms four leading FR VQAs (MS-SSIM, VQM, V-VIF, MOVIE) and one NR VQA (XT-PSNR).

3. DCT ANALYSIS OF COMPRESSED NATURAL VIDEO

The appearance of power laws in power spectral densities of natural scenes[12] suggests that it is reasonable to assume that there exist statistical relations between high-pass responses of natural images and their band-pass counterparts. Lossy video compression leads to distortion to the natural video, that usually manifests itself as loss of texture and other high frequency image features. Therefore, lossy compression will destroy the similarity between frequency feature maps of natural images. In the spatial domain, the distortion appears as an increase of the smooth image area and a decrease of the sharp image area.

Though an in-loop deblocking filtering technique is adopted to reduce blocking artifacts in H.264 compressed videos, blockiness sometimes remains visible as shown in Figure 1. A new blockiness metric is in need for H.264 compressed videos due to the failure of existing blockiness measurements. Our study shows blocking artifacts can be easily quantified based on the horizontal and vertical DCT components.

The behavior of the DCT coefficients of natural scenes have been intensively studied. The AC coefficients were conjectured to have Gaussian, Laplacian, or Cauchy distribution.[17] The lossy compression affects the natural distribution of the AC coefficients. In particular, coefficients equal to zero are much more frequent.

In summary, the lossy compression of videos with natural scenes leads to an increase of smooth image area, a decrease of sharp image area, blocking artifacts, a peaky AC coefficient distribution, and dissimilarity between frequency feature maps. To measure these types of distortion and predict the overall video quality, we apply the 4×4 DCT on a sliding window of size 4×4 in the decoded image. From the 16 DCT coefficients we derive six features per pixel per frame, each one yielding a feature map per frame.

Assume the frame size is $(M + 3) \times (N + 3)$. Only luminance is considered in our analysis, since the human visual system is more sensitive to luminance than chrominance. The generation of feature maps is as follows:

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}

Figure 2. Magnitudes of the DCT coefficients

Figure 3. An example of the decoded frame

- DCT map generation

  A sliding window of size 4×4 moves pixel by pixel over the entire frame of a decoded video. Fore each position the magnitudes C_1 to C_{16} of the corresponding DCT transform are computed, see Figure 2. This results in a DCT map with the size of $M \times N \times 16$ corresponding to $M \times N$ local windows and the magnitudes of 16 DCT coefficients of each window.

- Unsigned AC component feature map B_1

  Adding up the absolute values of the 15 AC coefficients of each block, we get the feature map B_1 of size $M \times N$. It is named the unsigned AC component feature map and will be used for the measurements of smoothness, sharpness and peakiness.

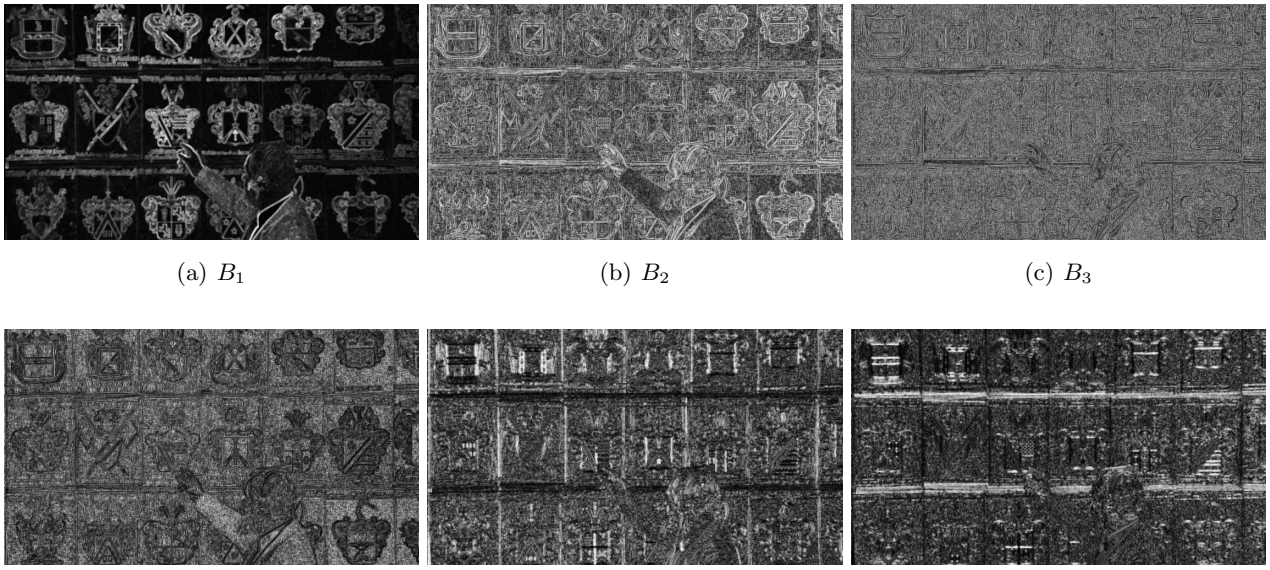
- AC coefficient normalization and feature maps B_2 to B_6

  For each block, dividing the unsigned AC coefficients by their sum for normalization, we get a matrix of normalized unsigned AC coefficients of size $M \times N \times 15$. From this matrix we obtain three frequency feature maps B_2, B_3, B_4, and two orientation feature maps B_5, B_6 to quantify dissimilarity and blockiness, respectively.

Table 1 indicates how the six feature maps are generated. An example of the decode frame is shown in Figure 3 and the six feature maps of the frame are illustrated in Figure 4. The feature map $B1$ contains all information of the image except DC components of local regions; B_2, B_3, B_4 contain low, medium, and high frequency components respectively; B_5 and B_6 contain vertical and horizontal components respectively. All of them have the same size of $M \times N$.

Table 1. Definition of feature maps B_1 to B_6

Feature map		name	description
Unsigned AC component		B_1	sum of $C_2, \cdots, C_{16}$
Frequency	low	B_2	sum of normalized C_2, C_5, C_6
	medium	B_3	sum of normalized $C_3, C_7, C_9, C_{10}, C_{11}$
	high	B_4	sum of normalized $C_4, C_8, C_{12}, C_{13}, C_{14}, C_{15}, C_{16}$
Orientation	vertical	B_5	sum of normalized C_2, C_3, C_4
	horizontal	B_6	sum of normalized C_5, C_9, C_{13}



(a) B_1 (b) B_2 (c) B_3

(d) B_4 (e) B_5 (f) B_6

Figure 4. Examples of feature maps B_1 to B_6 for the frame in Fig. 3

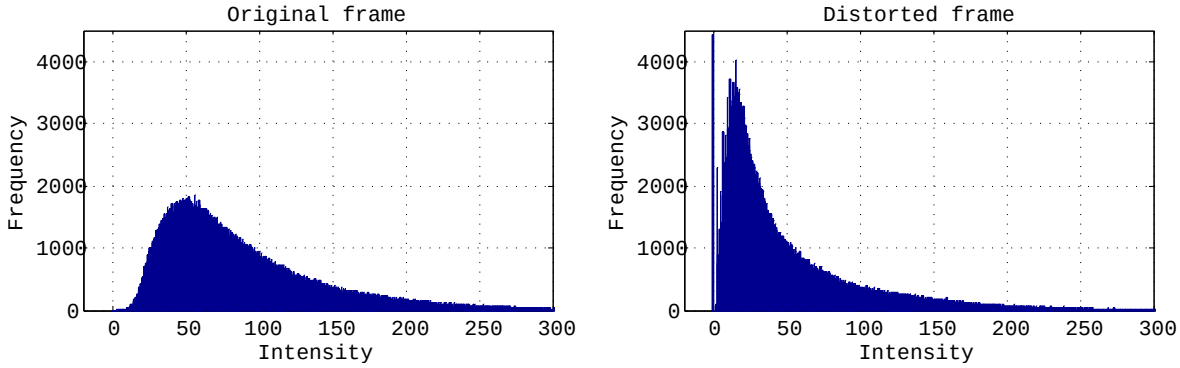


Figure 5. Histogram of the feature map B_1 for intensities up to 300 and bin size equal to 0.5

4. FRAME-LEVEL FEATURE EXTRACTION

Based on the six feature maps described in Section 3, five frame-level features, kurtosis, smoothness, sharpness, mean Jensen Shannon divergence (MJSD), and blockiness are extracted to quantify the distortion of compressed natural videos. The histograms of feature map B_1 extracted from an original frame and the corresponding distorted frame are illustrated in Figure 5. In contrast to the original frame, the distorted frame is observed to have a high histogram peak, a high frequency around zero, and low frequency at high intensities. The kurtosis, smoothness, and sharpness are designed to quantify the distortion based on the statistical properties of feature map B_1. The MJSD is designed to quantify the similarity between frequency feature maps based on the probability density functions of B_2, B_3, B_4. To quantify the blocking artifacts, the blockiness measurement is proposed based on B_5 and B_6. We list these features in Table 2.

Table 2. Features of frame t

$f_1(t)$	kurtosis	kurtosis of AC feature map B_1
$f_2(t)$	smoothness	relative smooth area of the current frame
$f_3(t)$	sharpness	relative edge area of the current frame
$f_4(t)$	MJSD	distribution distances of frequency feature maps (B_2, B_3) and (B_3, B_4)
$f_5(t)$	blockiness	measurement of blocking artifact caused by compression

4.1 Kurtosis

Let $p_1(x)$ be the probability density functions of B_1. The Kurtosis of $p_1(x)$ is chosen to measure the peakiness. It is given as:

$$f_1(t) = \text{Kurtosis} = \frac{E(x-\mu_x)^4}{\sigma_x^4} \in [1, \infty), \tag{1}$$

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

4.2 Smoothness

For each block, if the sum of AC coefficients is less than a threshold T_L, it is considered to be a smooth block. The degree of smoothness is defined as the relative area of the smooth region in the frame. The smoothness is expected to grow monotonically with the compression ratio.

$$f_2(x) = \text{Smoothness} = \frac{1}{MN} |\{(m,n)|B_1(m,n) < T_L\}| \in [0,1], \tag{2}$$

where $|A|$ denotes the number of elements of the set A.

4.3 Sharpness

For each block, if the sum of AC coefficients is greater than a threshold T_H, it is considered to be a sharp block. Sharpness is quantified as the relative area of the sharp region in the frame. A more compressed video is expected to have a smaller sharp area.

$$f_3(x) = \text{Sharpness} = \frac{1}{MN} |\{(m,n)|B_1(m,n) > T_H\}| \in [0,1]. \tag{3}$$

4.4 Mean Jensen Shannon divergence (MJSD)

The Jensen Shannon divergence (JSD) measures the "distance" between two probability distributions. Define $p(x)$ and $q(x)$ as two probability mass functions. Then the JSD is defined as the symmetrized version of Kullback-Leibler divergence (KLD), as follows:[18]

$$\text{KLD}(p||r) = \sum p(x) \log \frac{p(x)}{q(x)}, \tag{4}$$

$$\text{JSD}(p||q) = \frac{1}{2}(\text{KLD}(p||r) + \text{KLD}(q||r)), \tag{5}$$

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

It is observed that the similarity between two adjacent frequency maps of natural video is decreased due to lossy compression. To measure the decrease of their similarity, the mean JSD (MJSD) of B_2, B_3, and B_4 is defined as:

$$f_4(t) = \text{MJSD} = \frac{1}{2}\left(\text{JSD}(p_2||p_3) + \text{JSD}(p_3||p_4)\right) \in [0,1], \tag{6}$$

where $p_2(x), p_3(x)$, and $p_4(x)$ are the probability density functions of B_2, B_3, and B_4, respectively. A more compressed video is expected to have a larger MJSD.

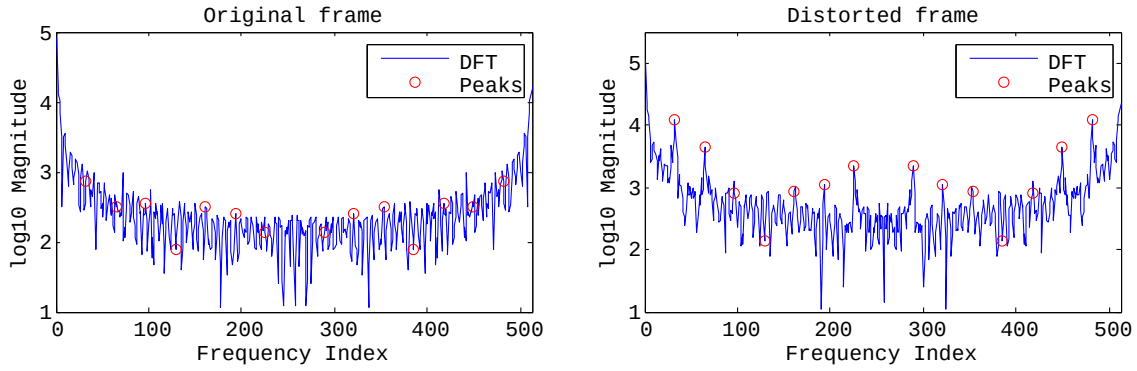


Figure 6. DFT coefficients for horizontal blockiness measurement

4.5 Blockiness

Due to the failure of existing blockiness metrics, we propose a new metric to measure blockiness in H.264 compressed natural videos. In our new metric, the blocking artifacts are measured based on the two orientation feature maps of the decoded frame, while the previous methods were based on the decoded frame itself. The horizontal and vertical blockiness are measured in the same way based on B_5 and B_6 respectively. The overall blockiness measurement is defined as the mean of horizontal and vertical blockiness measurements.

Assume the macroblock size is known as $s \times s$. We measure the horizontal blockiness by applying a sum operation along each row in feature map B_6. It results in a 1-D array of length M, denoted as P_{H}.

$$P_{\mathrm{H}}(m) = \sum_{n=0}^{N-1} B_6(m,n), \quad m = 0, ..., M-1. \tag{7}$$

It is difficult to distinguish directly the blockiness power from P_{H}. Fortunately, more clues can be obtained when we go to the frequency domain.[7] We take the 1-D DFT of P_{H} and consider the magnitude of the DFT coefficients. These are given by

$$F_{\mathrm{PH}}(l) = \left| \sum_{m=0}^{M-1} P_{\mathrm{H}}(m) \exp\left(-\frac{j2\pi ml}{L}\right) \right|, \tag{8}$$

where $l = 0, ..., L-1$ and L is the smallest power of two greater or equal to M.

Due to the nature of the DFT, $F_{\mathrm{PH}}(l)$ has peaks at $l = \frac{L}{s} \cdot i$, for $i = 1, 2, ..., \frac{s}{2} - 1$. Values at those peaks are closely related to the horizontal blockiness of the image. The horizontal blockiness measurement is then computed as:

$$B_{\mathrm{MH}} = \frac{1}{\frac{s}{2} - 1} \sum_{i=1}^{s/2-1} \log_{10}\left(F_{\mathrm{PH}}\left(\frac{L}{s} \cdot i\right) + 1\right) \in [0, \infty). \tag{9}$$

Figure 6 illustrates the DFT coefficients of a reference frame and the corresponding distorted frame. A 512-point DFT was taken and the macroblock size is 16×16. No periodic peak is observed in the left subfigure, while periodic peaks appear at $l = 32, 64, 96, 128, 160, 192, 224$ in the right subfigure. Values at these peaks are chosen for computing B_{MH} in Equation 9. Note that due to the symmetry of the DFT, 14 peaks rather than 7 peaks are marked in each subfigure, but only the first 7 peaks are used in computation.

Applying a sum operation along each column in feature map B_5, the vertical blockiness B_{MV} is then measured in the same way. Finally, the overall blockiness measurement is computed by

$$f_5(t) = \text{Blockiness} = \frac{B_{\mathrm{MH}} + B_{\mathrm{MV}}}{2} \in [0, \infty). \tag{10}$$

Higher values associate with more compressed videos.

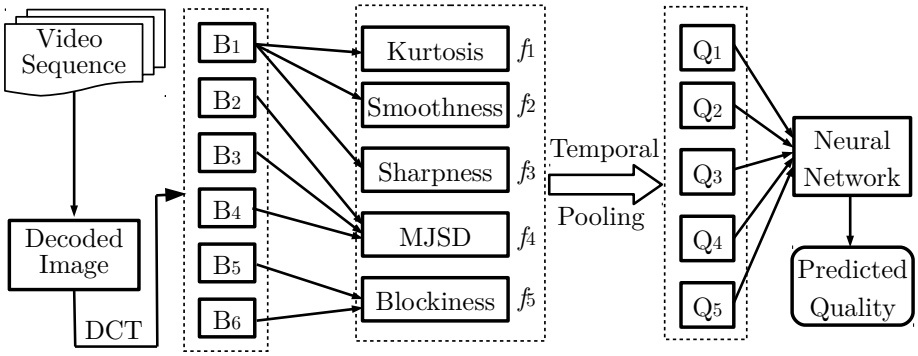


Figure 7. Flowchart of the proposed model DVQPM

5. PREDICTION MODEL

To predict the video quality from the frame-level features of all frames in one video sequence, we design a prediction model which is composed of temporal pooling and a multilayer neural network. Each frame-level feature yields a vector $(f_j(1), f_j(2), \cdots, f_j(t), \cdots, f_j(T_0))$ and is transformed to a video-level feature by Minkowski pooling strategy which is given as follows:[19]

$$Q_j = \sqrt[4]{\frac{1}{T_0}\sum_{t=1}^{T_0} f_j(t)^4}, \tag{11}$$

where $j = 1, 2, \cdots, 5$, and T_0 is the number of frames in the video sequence.

The video-level features $Q_1, \ldots, Q_5$ are then treated as inputs to a neural network trained to predict the subjective video quality score. In our implementation, the neural network was comprised of 10 hidden nodes. The predicted quality given by the trained neural network should be as close as possible to the subjective assessment.

Figure 7 gives the high level organization of the proposed prediction architecture DVQPM. It is composed of three stages. In the first stage, six feature maps are generated from the DCT coefficients as listed in Table 1. Second is a frame-level feature extraction stage, as described in Section 4 and Table 2. In the last stage, each extracted frame-level feature as a vector is first taken as input to the temporal pooling. A single score is yielded as the corresponding video-level feature along the time axis. An objective video quality score is then predicted by the trained neural network from video-level features.

6. PERFORMANCE EVALUATIONS

We evaluate the performance of the proposed method on the popular LIVE video database.[20,21] It consists 10 reference videos and 150 distorted videos with resolution of 768×432 pixels and length of 10 seconds. A total of 15 test sequences were generated from each of the reference sequences using four different distortion processes: MPEG-2 compression (4 test videos per reference), H.264 compression (4 test videos per reference), lossy transmission of H.264 compressed bitstreams through simulated IP networks (3 test videos per reference) and lossy transmission of H.264 compressed bitstreams through simulated wireless networks (4 test videos per reference). Each distorted videos was evaluated by 38 human observers. The subjective evaluation was performed using a single stimulus paradigm with hidden reference removal,[22] where the observers were asked to provide their opinion of video quality on a continuous scale. The mean opinion score (MOS) in the range of $[0, 100]$ is provided as the subjective quality assessment of each distorted video.

Since our method aims to objectively assess the quality of compressed videos and the most popular compression model is H.264, only the 40 H.264 compressed videos in the database were used to evaluate the performance. We set $T_L = 1$ in Eq. 2 and $T_H = 300$ in Eq. 3. The leave-one-out strategy was adopted for training and testing,

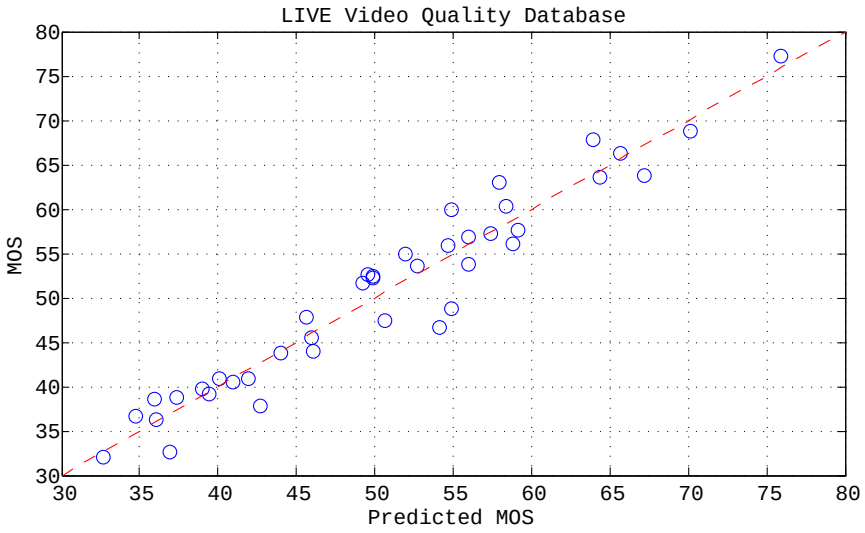


Figure 8. Scatter plot of predicted MOS vs. MOS

thus 39 video sequences were chosen for neural network training and validation; the remaining sequence was for testing. The training and testing process was performed 40 times with a different test video sequence each time. Let array $\mathbf{a} = (a_1, a_2, \cdots, a_{40})$ be the predicted quality of the proposed DVQPM, and $\mathbf{b} = (b_1, b_2, \cdots, b_{40})$ be the corresponding MOS of the subjective quality assessment. Figure 8 is the scatter plot of the two arrays. It indicates a good match between the predicted objective quality and the subjective quality.

Four statistical indexes were used to evaluate the performance. They are the Linear (Pearson's) Correlation Coefficient (LCC), the Spearman's Rank Ordered Correlation Coefficient (SROCC), the Root Mean Squared Error (RMSE), and the Mean Absolute Error (MAE) between the predicted scores $\mathbf{a}$ and the mean opinion scores (MOS) $\mathbf{b}$. A value close to 1 for SROCC and LCC and a value close to 0 for RMSE and MAE indicates superior correlation with subjective assessment. The four indexes are defined as follows:[23,24]

$$\mathrm{LCC}(\mathbf{a},\mathbf{b}) = \frac{\sum_{k=1}^{K}(a_k-\bar{a})(b_k-\bar{b})}{\sqrt{\sum_{k=1}^{K}(a_k-\bar{a})^2}\sqrt{\sum_{k=1}^{K}(b_k-\bar{b})^2}}, \tag{12}$$

$$\mathrm{SROCC}(\mathbf{a},\mathbf{b}) = \frac{\sum_{k=1}^{K}(u_k-\bar{u})(v_k-\bar{v})}{\sqrt{\sum_{k=1}^{K}(u_k-\bar{u})^2}\sqrt{\sum_{k=1}^{K}(v_k-\bar{v})^2}}, \tag{13}$$

$$\mathrm{RMSE}(\mathbf{a},\mathbf{b}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(a_k-b_k)^2}, \tag{14}$$

$$\mathrm{MAE}(\mathbf{a},\mathbf{b}) = \frac{1}{K}\sum_{k=1}^{K}|a_k-b_k|, \tag{15}$$

where $\bar{a}$ and $\bar{b}$ are the mean values of $\mathbf{a}$ and $\mathbf{b}$, respectively, u_k is the rank of a_k in array $\mathbf{a}$, and v_k is the rank of b_k in array $\mathbf{b}$. The performance of DVQPM in terms of the four indexes in the LIVE video database is tabulated in Table 3.

We compared the performance of our model with four FR VQA algorithms and one NR VQA in terms of SROCC and LCC in Table 4. In the table, the performances of MS-SSIM,[1] VQM,[4] V-VIF,[5] and MOVIE[6] were evaluated based on the H.264 videos in the LIVE Video Database, and the results were provided in Refs. 20, 21. For XT-PSNR, the authors created their own H.264 video database, assessed the subjective quality of all videos,

Table 3. Performance of DVQPM

Database	LIVE video database
LCC	0.9666
SROCC	0.9630
RMSE	2.7807
MAE	2.1833

Table 4. Performance Comparison

Indices	MS-SSIM	VQM	V-VIF	MOVIE	XT-PSNR	DVQPM
LCC	0.6919	0.6459	0.6911	0.7902	0.9530	**0.9666**
SROCC	0.7051	0.6520	0.6807	0.7664	0.9460	**0.9630**

and evaluated the performance on their database rather than the LIVE video database. The evaluation results of XT-PSNR listed in Table 4 were provided in Ref. 16.

It is clear that DVQPM outperforms other algorithms in terms of LCC and SROCC. Although our model is competitive with the four "general-purpose" FR VQA algorithms, the proposed metric is distortion-specific. Its performance for other types of distortion has not been evaluated yet. With higher SROCC and LCC, our model shows better performance than XT-PSNR which was also proposed for H.264 distorted videos. Note that RMSE and MAE were also provided in Ref. 16, but they are not comparable for they are sensitive to the scale of the MOS, and a different MOS scale was used in their database when subjective assessment was performed.

7. CONCLUSION

To assess the objective quality of H.264 coded videos instead of expensive and time-consuming subjective assessment, a NR VQA model named DVQPM is proposed based on local DCT coefficients. Studying the properties of natural scenes and types of distortion in compressed videos, DVQPM combined the existing artifact-based and NNS-based approaches and outperformed the leading VQA methods according to the evaluation results in the LIVE video database.

The proposed DVQPM quantifies the distortion of a video sequence frame by frame. A DCT is 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 feature maps are extracted from the DCT map. Five frame-level features, including three artifacts metrics (smoothness, sharpness and blockiness) and two statistical metrics (kurtosis and MJSD), are extracted from these feature maps. The frame-level features are transformed to video-level features through temporal pooling. Finally a trained multilayer neural network provides the predicted video quality according to the five video-level features.

The performance evaluation was conducted on the LIVE video database. The results for 40 H.264 coded videos show a strong correlation between the predicted quality and subjective quality. It's also clear that DVQPM outperforms other four leading FR VQA algorithms and one NR VQA algorithm in terms of LCC and SROCC.

The proposed DVQPM is designed for H.264 coded videos to evaluate the performance of imaging systems based on H.264 compression standard, such as mobile phone cameras, HD camcorders, and video surveillance systems. Its application is limited to H.264 coded videos. In our future work, to extract more efficient features, we will further study the properties of natural scenes and the influence of various compressions on these properties. The blockiness metric proposed in the paper is used as a frame-level feature to measure the blocking artifact of H.264 coded videos. It can be extended to measure the blockiness in JPEG and JPEG 2000 compressed images or MPEG-2 compressed videos.

REFERENCES

[1] Wang, Z., Simoncelli, E. P., and Bovik, A. C., "Multi-scale structural similarity for image quality assessment," in [*IEEE Asilomar Conference on Signals, Systems and Computers*], (Nov. 2003).

[2] Wang, Z., Lu, L., and Bovik, A. C., "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication* **19**, 121–132 (Feb. 2004).

[3] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing* **13**, 600–612 (Apr. 2004).

[4] Pinson, M. H. and Wolf, S., "A new standardized method for objectively measuring video quality," *IEEE Transactions Broadcasting* **50**, 312 – 322 (Sept. 2004).

[5] Sheikh, H. R. and Bovik, A. C., "A visual information fidelity approach to video quality assessment," in [*The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*], 23–25 (Nov. 2005).

[6] Seshadrinathan, K. and Bovik, A. C., "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing* **19**, 335 – 350 (Feb. 2010).

[7] Wang, Z., Bovik, A. C., and Evans, B. L., "Blind measurement of blocking artifacts in images," in [*International Conference on Image Processing*], **3**, 981–984 (2000).

[8] Bailey, D., Carli, M., Farias, M., and Mitra, S., "Quality assessment for block-based compressed images and videos with regard to blockiness artifacts," in [*Tyrrhenian International Workshop on Digital Communications*], (2002).

[9] Liu, H. and Heynderickx, I., "A no-reference perceptual blockiness metric," in [*International Conference on Acoustics, Speech, and Signal Processing*], 865 – 868 (2008).

[10] Chen, C. and Bloom, J. A., "A blind reference-free blockiness measure," in [*the 11th Pacific Rim conference on Advances in multimedia information processing*], 112–123 (2010).

[11] Simoncelli, E. P. and Olshausen, B. A., "Natural image statistics and neural representation," *Annual Review of Neuroscience* **24**, 1193–1216 (May 2001).

[12] Geisler, W. S., "Visual perception and the statistical properties of natural scenes," *Annual Review of Psychology* **59**, 167–192 (Aug. 2007).

[13] Olshausen, B. A. and Field, D. J., "Natural image statistics and efficient coding," *Network: Computation in Neural Systems* **7**, 333–339 (Jan. 1996).

[14] Geisler, W. S. and Ringach, D., "Natural system analysis," *Visual Neuroscience* **26**, 1–3 (Jan. 2009).

[15] Moorthy, A. K. and Bovik, A. C., "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing* **20**, 3350–3364 (Dec. 2011).

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

[17] Altunbasak, Y. and Kamaci, N., "An analysis of the DCT coefficient distribution with the H.264 video coder," in [*IEEE International Conference on Acoustics, Speech and Signal Processing*], **3**, 177–180 (May 2004).

[18] Cover, T. M. and Thomas, J. A., [*Elements of Information Theory*], John Wiley & Sons, second ed. (2006).

[19] Wang, Z. and Bovik, A. C., [*Modern Image Quality Assessment*], Morgan & Claypool Publishers, first ed. (2006).

[20] Seshadrinathan, K., Soundararajan, R., Bovik, A. C., and Cormack, L. K., "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing* **19**, 1427–1441 (June 2010).

[21] Seshadrinathan, K., Soundararajan, R., Bovik, A. C., and Cormack, L. K., "A subjective study to evaluate video quality assessment algorithms," in [*SPIE Proceedings Human Vision and Electronic Imaging*], (Jan. 2010).

[22] Pinson, M. and Wolf, S., "Comparing subjective video quality testing methodologies," in [*SPIE Video Communications and Image Processing Conference*], 8–11 (2003).

[23] Myers, J. L. and Well, A. D., [*Research Design and Statistical Analysis*], Routledge, second ed. (2002).

[24] Cui, L. and Allen, A. R., "An image quality metric based on corner, edge and symmetry maps," in [*British Machine Vision Conference*], (2008).