2013 IEEE International Conference on Image Processing : ICIP 2013, Proceedings ; September 15-18, 2013, Melbourne, Victoria, Australia / IEEE Signal Processing Society. - Piscataway, NJ : IEEE Service Center, 2013. - S. 49-53. - ISBN 978-1-4799-2341-0

A NO-REFERENCE VIDEO QUALITY ASSESSMENT BASED ON LAPLACIAN PYRAMIDS

Kongfeng Zhu[⋆] *Keigo Hirakawa*[†] *Vijayan Asari*[†] *Dietmar Saupe*[⋆]

[⋆] Department of Computer and Information Science, University of Konstanz, Germany
[†] Department of Electrical and Computer Engineering, University of Dayton, USA

ABSTRACT

This paper presents an approach to predict the quality of compressed videos with content of natural scenes. The method is focused on measuring the distortion of compressed video without reference. There are two main steps of the proposed method: measuring distortion and predicting video quality. Each frame of the distorted video sequence is first decomposed to an N-subband Laplacian pyramid, then their intra-subband and inter-subband statistical features are fully exploited. Three intra-subband features and three inter-subband features are taken as inputs of the prediction model. Its output is a single score as the predicted video quality. The performance of the proposed method is evaluated on the LIVE video database and the LIVE mobile video database. Results show that the predicted quality scores are well correlated with the mean opinion scores associated to the subjective assessment.

Index Terms— Image/video quality assessment, no-reference, natural scenes, Laplacian pyramid

1. INTRODUCTION

The increasing demands and the limits on bandwidth for image/video capture, transmission, and storage lead to occurence of information loss and extraneous artifacts. How the distortion affects the quality of viewing experience has become the interest of researchers in visual quality assessment. Naturally, the subjective assessment is the golden standard, yet it is time-consuming, cumbersome and impractical. Hence one seeks to develop algorithms such that the predicted quality of distorted visual stimuli has high correlation with subjective assessment.

Depending on the amount of available information, objective quality assessment can be divided into three categories: full-reference (FR), reduced reference (RR), and no-reference (NR) approaches. Without *a priori* knowledge about the pristine image or video, NR image/video quality assessment (IQA/VQA) is the most useful but also most difficult one to accurately predict visual quality.

A great effort has been made for NR IQA based on the statistics of natural images [1, 2, 3]. Undistorted natural images are considered to possess certain statistical properties that hold across different image contents. The natural scene statistics (NSS) models seek to capture those statistical properties of natural scenes that 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. In this paper, any image or video obtained from a camera or camcorder is considered to be natural [2, 4].

The Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index is a NR IQA algorithm that popularized image quality assessment based on NSS [5]. 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, a no-reference quality assessment metric was proposed for digital video subject to H.264/AVC encoding [6]. Assuming that DCT coefficients are corrupted by quantization noise, the coding error was estimated first, then a spatio-temporal contrast sensitivity function applied to the DCT domain to perceptually weight the estimated coding error.

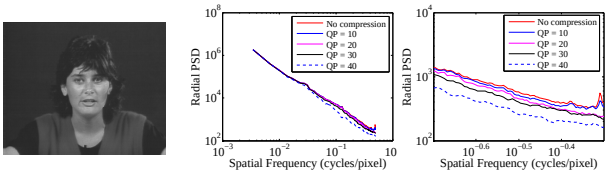
Inspired by previously proposed models for NR IQA and VQA, in this paper we designed an NR IQA based on the NSS model to predict the quality of encoded video sequences frame by frame. In the proposed method, many fewer features were extracted than for that in [5] and the reference images used in [6] for training are not needed. Moreover, we adopted a two-layer neural network that simplified the training procedures in [5, 6]. Our NSS-IQA model performs very well on video databases, and its performance on image databases will be evaluated in our future work.

The rest of this paper is organized as follows. In Section 2 we investigate the statistical properties of image subband decompositions, paying a particular attention to the inter-subband dependencies. In Section 3, we describe the model of the proposed method and how these features are used for distortion measurement as well as quality prediction. The performance of the proposed approach is evaluated in Section 4 and the paper is concluded in Section 5.

2. DISTORTION MEASUREMENT

2.1. Frequency analysis of compressed natural video

Lossy video compression leads to unavoidable distortion to natural video, which usually manifests itself as loss of texture and other high frequency image features [7]. This can



(a) Miss America (b) PSD (c) PSD closeup

Fig. 1. Power spectral density of compressed natural video

be seen by comparing the radial power spectral density (PSD) of a distorted object in the compressed video with that of the distortion-free object to quantify preserved texture. Consider the radial power spectral density (PSF) of "Miss America" video, shown in Figure 1. The closeup of the reference video and its four compressed versions with different H.264 quantization parameters (QP) suggests that PSD curves are overlapped at low frequency, but divergent for higher frequencies.

We may conclude from Figure 1 that the relative decay of power in spatial frequency decompositions is a useful predictor of compressed image quality. Indeed, this is relatively easy to do in the FR context where the true rate of PSD decay can be measured. The main challenge of developing NR IQA is to detect changes in the statistics of high frequency image features without direct access to the reference.

In practice, canonical frequency decomposition of image and video signals such as the Fourier transform ignore spatial locality. For this reason, we employ the expanded Laplacian pyramid as a compromise between frequency and spatial decomposition [8]. Let $I(i,j)$ be an image or a video frame function represented in the spatial domain, where $(i,j) \in \mathbb{Z}^2$ is the pixel index. Then we denote by

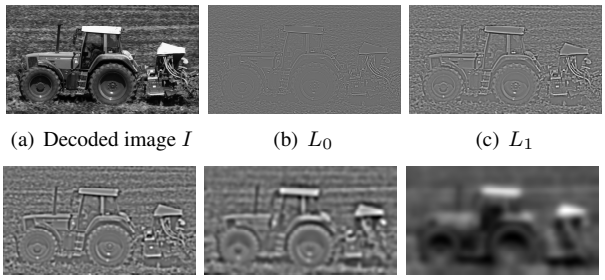
$$I \mapsto \{L_0, L_1, \ldots, L_{N-1}\} \tag{1}$$

the expanded Laplacian pyramid decomposition of image $I(i,j)$ represented as a series of quasi-bandpassed images. Each subband image $L_n(i,j)$ is the size of the original image $I(i.j)$. As shown by Figure 2, the fine details are captured in $L_0(i,j)$ while progressively coarser features are prominent in higher levels.

With access to high frequency image details via subband decomposition, one can assess the distortion caused by lossy compression by extracting its statistical features. To meet the objective to assess the relative decay of power in spatial frequency decomposition, we propose two approaches to model the Laplacian pyramid coefficients. First is the computation of intra-subband statistics, which will be compared across subbands to gauge the rate of decay (that is influenced by compression). Second is the incorporation of inter-subband features aimed at directly assessing the degree of persistence across scale [9].

2.2. Intra-subband features

Intra-subband statistics that reflect statistical properties of one single subband can be useful for predicting the rate of decay



(a) Decoded image I (b) L_0 (c) L_1

(d) L_2 (e) L_3 (f) L_4

Fig. 2. One frame of a video sequence and 5 subbands of the corresponding expanded Laplacian pyramid

across frequency. In our study, mean, variance, skew, energy, entropy, and kurtosis of empirical Laplacian pyramid coefficients were considered. In particular, we found that energy, entropy, and kurtosis of coarse scale subbands were less subject to compression distortions compared to the finer scale subbands (see details in Section 2.4). Moreover we claim that the fine-scale-to-coarse-scale ratio of subband statistics is a good proxy for the rate of decay across spatial frequency that is stable for all frames in compressed videos. Hence we conclude that the ratio of intra-subband statistics offers a stable prediction of degradation due to compression.

Let us view the subband coefficients in L_n as stationary random processes, where the random variable $X = L_n(i,j)$ have the same probability mass function $p(x)$. Then the statistics considered in this article are defined as follows:

$$E_n = \log_{10}\left(\Sigma_i \Sigma_j L_n^2(i,j)\right), \tag{2}$$

$$H_n(X) = -\Sigma p(x) \log p(x), \tag{3}$$

$$\kappa_n(x) = E(x-\mu_x)^4/\sigma_x^4. \tag{4}$$

2.3. Inter-subband features

Apparent power laws in the PSD [2] and persistence across scale in space-frequency decompositions [8] suggest that it is reasonable to assume that there exist statistical relations between high-pass responses of natural images and their bandpass counterparts. Indeed, in our studies, we found that such a relationship exists for natural images and this relationship is distorted by compression. To quantify the persistence across scale, we consider the following inter-level features.

1) Jensen Shannon divergence (JSD) is a measure of the "distance" between two probability distributions which can be generalized to measure the distance (dissimilarity) between a finite number of distributions. Define $p(x)$ and $q(x)$ as two probability mass functions of two images. Then the JSD is defined as the symmetrized version of Kullback-Leibler divergence (KLD), as follows [10]:

$$\text{KLD}(p||q) = \sum p(x) \log(p(x)/q(x)) \tag{5}$$

$$\text{JSD}(p||q) = \frac{1}{2}(\text{KLD}(p||r) + \text{KLD}(q||r)) \tag{6}$$

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

2) The Structural Similarity Index (SSIM) is a popular FR IQA based on luminance/contrast/structure similarities [11]. It is given by

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (7)$$

Refer to [11] for detailed explanation on the above equation. In the context of assessing persistence across scale, we use the mean of SSIM index map as a way to quantify the dependency between two Laplacian pyramid subbands L_n and L_m, denoted as $\text{MSSIM}(L_n, L_m)$.

3) Smoothness is a measurement of the relative size of flat area in an image. Although absolutely flat regions do not occur in natural scenes, they exist in compressed images and videos due to aggressive quantization. To identify flat regions based on the Lapalacian pyramid, the SSIM index map [11] between I and L_{N-1} is computed, and denoted as S_{SSIM}. When the SSIM index is greater than T_0, the local region is considered to be a flat region. Thus, the flat region is given by

$$S_{\text{smooth}}(i,j) = \begin{cases} 1, & \text{if } S_{\text{SSIM}}(i,j) > T_0; \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

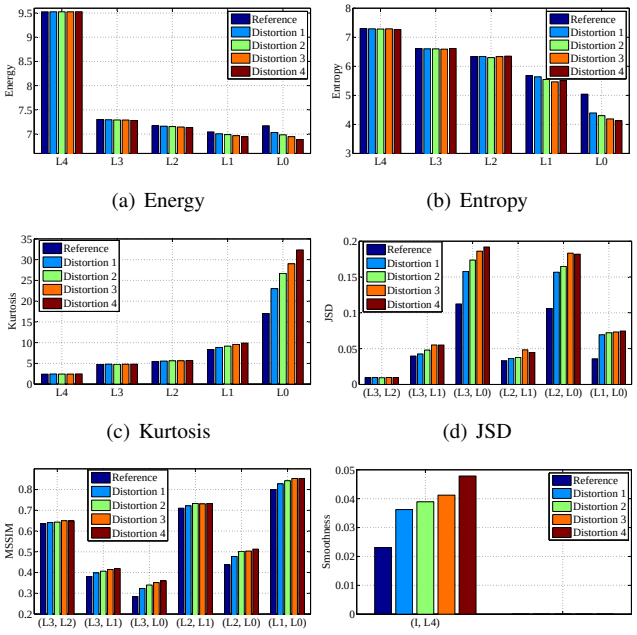
The smoothness of the frame is defined as the mean of S_{smooth}. T_0 is set to 0.95, and the size of the local region in the computation of $S_{\text{SSIM}}(i,j)$ is set to 9×9 in the experiment in Section 4.

2.4. Selection of frame features

To choose the most efficient features, we further studied the intra- and inter-subband features. In Figure 3, (a-c) show intra-subband features and (d-f) show inter-subband features of the same frame in the reference video and four H.264 distorted videos, respectively. 'Reference' stands for the distortion-free video. 'Distortion 1' has the lowest compression ratio while 'Distortion 4' has the highest compression ratio. The frame is decomposed in a 5-subband Laplacian pyramid, i.e. L_0, L_1, L_2, L_3, L_4.

It is found that the intra-subband features of L_0, the JSD and MSSIM between L_0 and the coarsest bandpass L_{N-2}, and smoothness are most sensitive to H.264 compression. It is observed that the energy and entropy of L_0 decrease, and kurtosis increases with respect to compression ratio, while the corresponding features of L_2, L_3, and L_4 stay roughly constant. JSD and MSSIM generally increase with respect to compression ratio, but JSD between the distribution in L_0 and L_3, and MSSIM between L_0 and L_3 change dramatically with respect to compression ratio. The same relationship is observed between smoothness and compression ratio.

For the sake of computational efficiency, only the intra- and inter-subband features of L_0 and L_3 are considered in our prediction model. To normalize intra-subband features, the energy and entropy of L_0 are divided by the corresponding features of L_3, and the kurtosis of L_3 is divided by kurtosis of L_0. The selected features of frame t, $f_l(t), l = 1, ...6$, in a video sequence are listed in Table 1.



(a) Energy (b) Entropy

(c) Kurtosis (d) JSD

(e) MSSIM (f) Smoothness

Fig. 3. Intra- and inter-subband features

Table 1. Features of frame t

$f_1(t)$	E_0/E_3	Energy ratio of L_0 and L_3
$f_2(t)$	H_0/H_3	Entropy ratio of L_0 and L_3
$f_3(t)$	κ_3/κ_0	Kurtosis ratio of L_3 and L_0
$f_4(t)$	$\text{JSD}(L_0, L_3)$	JSD between L_0 and L_3
$f_5(t)$	$\text{MSSIM}(L_0, L_3)$	MSSIM between L_0 and L_3
$f_6(t)$	Smoothness	Relative size of flat area

3. PREDICTION MODEL

To predict the video quality from the extracted intra- and inter-subband features, a prediction model is designed to pool all features into one single score, as the predicted quality of input video sequence. Thus, for one video sequence, the inputs of the prediction model are the six frame features $f_1(t), ..., f_6(t)$ of all frames, and the output is a single value which is the predicted video quality. The predicted quality should be as close as possible to the subjective assessment.

Figure 4 gives the high level organization of the proposed prediction architecture. It is composed of two stages. First is a feature extraction stage, as described in Table 1. The computed statistics are taken as inputs to the temporal pooling using 4th-order Minkowski norm as [12] suggested to yield a single score as the corresponding video feature along the time axis:

$$Q_{f_l} = \sqrt[4]{\sum_t f_l(t)^4}, \quad (9)$$

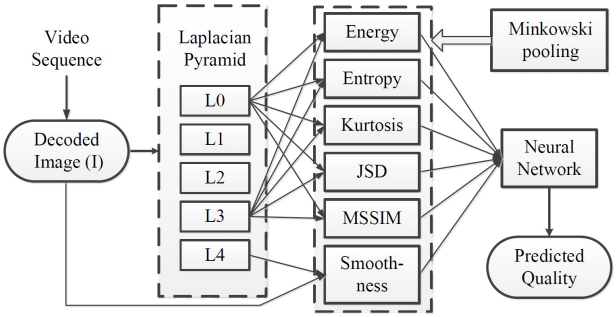


Fig. 4. Flowchart of the proposed method

where $l = 1, 2, ..., 6$. We assume that all video sequences have the same length, otherwise, normalization is needed. In the second stage, the scores $Q_{f1}, \ldots, Q_{f6}$ are treated as input to a neural network trained to predict the subjective video quality score. In our implementation, the neural network was comprised of 20 hidden nodes.

4. PERFORMANCE EVALUATION

We evaluated the performance of the proposed method on the LIVE video quality database (LIVE VQDB) [13, 14] and LIVE mobile video quality database (LIVE MVQDB) [15]. The LIVE VQDB consists of 10 reference videos and 150 distorted videos with resolution of 768×432 pixels and length of 10 seconds. The mean opinion score (MOS) in the range of $[0, 100]$ is provided as the subjective quality assessment of each distorted video. The LIVE MVQDB consists of 10 raw HD reference videos and 200 distorted videos, each of resolution 1280×720 at a frame rate of 30 fps, and of duration 15 seconds each. Here, the MOS of each video is in the range $[0, 5]$.

Since our method aims to objectively assess the quality of compressed videos and the most popular compression model is H.264, only the H.264 compressed videos in each database were used to evaluate the performance. Each of the two databases contains 10 sets of H.264 compressed videos. Each set contains four videos generated from one reference video. The leave-one-out strategy was adopted for training and testing, thus 9 sets of distorted video sequences (36 videos in total) were chosen for neural network training and validation; the remaining distorted video set was for testing. The training and testing process was performed 10 times with a different set of distorted videos each time.

Four indices 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 and the mean opinion scores (MOS) obtained from the subjective assessment. 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 scatter plots of objective VQA scores vs. MOS for H.264 videos in LIVE VQDB and LIVE MVQDB are illustrated in Figure 5. The corresponding SROCC, LCC, RMSE and MAE values are tabulated in Table 2.

Table 2. Performance of the proposed metric

Database	LCC	SROCC	RMSE	MAE
LIVE VQDB	0.9112	0.9396	4.4884	3.7670
LIVE MVQDB	0.9353	0.9281	0.4041	0.3334

Table 3. Performance Comparison on LIVE VQDB

Indices	MS-SSIM	V-VIF	MOVIE	Proposed
LCC	0.6919	0.6911	0.7902	**0.9112**
SROCC	0.7051	0.6807	0.7664	**0.9396**

The performance comparison in Table 3 shows our NR VQA is competitive with the "general-purpose" FR VQA algorithms (MS-SSIM [16], V-VIF [17], MOVIE [18]) listed in [14] for H.264 coded videos. However, the proposed metric is distortion-specific. The state-of-the-art algorithm proposed by Brandão and Queluz [6] is the only one we found so far with the same purpose as our algorithm, but it is not comparable since a different database was used for evaluation in [6].

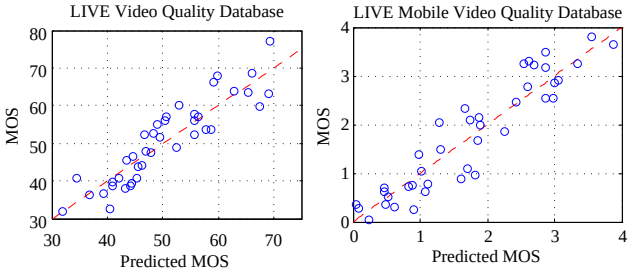


Fig. 5. Scatter plots of predicted MOS vs. MOS

5. CONCLUSION

An NSS-based NR IQA algorithm is proposed in this paper to evaluate the quality of compressed videos with content of natural scenes frame by frame. The method aims to measure the distortion of fine details based on Laplacian pyramids, since the distortion introduced by lossy compression mainly appears at high frequency.

Evaluation results on two video databases show that the predicted quality scores are well correlated with the MOS. The proposed NR method is competitive with the widely accepted FR VQA algorithms in predicting the quality of compressed videos. However it is distortion-specific, while FR VQA methods are designed for general purpose. In future, the method will be improved to a general-purpose NR IQA/VQA method to predict the quality of images or videos with various kinds of distortion.

6. REFERENCES

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

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

[3] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," vol. 7, pp. 333–339, Jan. 1996.

[4] W. S. Geisler and D. Ringach, "Natural system analysis," *Visual Neuroscience*, vol. 26, pp. 1–3, Jan. 2009.

[5] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

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

[7] K. Zhu, S. Li, and D. Saupe, "An objective method of measuring texture preservation for camcorder performance evaluation," in *IS&T SPIE Electronic Imaging 2012*, vol. 8293, no. 6, Burlingame, CA, USA, Jan. 2012.

[8] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, pp. 532–540, Apr. 1983.

[9] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 886–902, 1998.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

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

[12] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, 1st ed. Morgan & Claypool Publishers, 2006.

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

[14] ——, "A subjective study to evaluate video quality assessment algorithms," in *SPIE Proceedings Human Vision and Electronic Imaging*, Jan. 2010.

[15] A. K. Moorthy, L. K. Choi, G. de Veciana, and A. C. Bovik, "Subjective analysis of video quality on mobile devices," in *Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, Arizona, Jan. 2012.

[16] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, Nov. 2003.

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

[18] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335 – 350, Feb. 2010.