EFFICIENT FULL-REFERENCE ASSESSMENT OF IMAGE AND VIDEO QUALITY

Patrick Ndjiki-Nya, Mikel Barrado, and Thomas Wiegand

Image Communication Group, Image Processing Department
Fraunhofer Institute for Telecommunications - Heinrich-Hertz-Institut, Berlin, Germany
{ndjiki/wiegand}@hhi.de

ABSTRACT

Automatic quality assessment of digital pictures is a crucial issue in several image and video processing applications including broadcasting, archiving, or restoration. The visibility of impairments related to digital image processing systems is subject to the spatio-temporal properties of the given image or video content. As quality assessment using subjective tests carried out by humans is very costly, time consuming, and not compatible with real-time constraints, objective measures are required that can predict the perceptual judgment of human viewers. In this paper, a top-down quality assessment tool that mimics a selection of prominent human visual system properties is presented. Quality evaluation is conducted based on perceptually salient feature points in our approach. It is shown that the proposed method, although featuring a significantly lower complexity than standardized quality measures, performs as well as these for block-based hybrid video coding.

Index Terms— Image, video, quality assessment, block transform coding

1. INTRODUCTION

Many digital video applications require dynamic monitoring and adjustment of image quality. For that, low complexity objective measures are required that can predict the perceptual judgment of the Human Visual System (HVS). Hence, major efforts have been made by the Video Quality Experts Group [1] (VQEG) to establish standardized quality models.

So-called full-reference measures assume both, original and distorted signals to be known. Many of these measures are based on the evaluation of the difference signal between the reference and the processed image signal [2] (e.g. MSE, PSNR). Although they poorly correlate with human visual perception [3], these measures are very widely used due to their small complexity. Visibility-based approaches achieve some improvements by considering the visibility of differences between the original and processed signal based on HVS properties [4]. The accuracy of such bottom-up measures is typically limited as knowledge of the HVS is still in its infancy. In opposition to bottom-up models, top-down approaches have been described in the literature [5],[6]. They typically do not carry out a pixel-wise comparison of the distorted and reference signal. These approaches either compare global disparities between the two signals and thereby usually mimic a selection of assumed HVS functionalities.

In this work, a spatial full-reference quality measure (QM) is presented. The QM is optimized in terms of its efficiency and its complexity. In the remainder of the paper, an in-depth description of the proposed QM is given (Sec. 2). The latter is evaluated for block-based hybrid video coding. The corresponding experimental results are presented in Sec. 3, where the performance evaluation criteria are also introduced.

2. PROPOSED QUALITY MEASURE

2.1 Principle and Artifacts of Block Transform VC

Standardized block-based hybrid video coding (VC) algorithms such as H.264/AVC process an input picture macroblock-wise [7]. The latter are coded either using intra or inter prediction modes. For macroblocks coded with intra mode, a given prediction macroblock is determined using only information from the currently processed picture, i.e. from spatially neighboring samples. In the inter mode, a given prediction macroblock is determined through motion-compensated prediction using one or more previously decoded reference pictures [7]. The artifacts that typically occur in the given framework are blocking (tiling), blurring as well as unnatural, jerky motion.

In this work, the spatial VQM defined by Ong et al. [8],[9] is revisited with the aims of simplification and significant performance improvement. It is believed that a much simpler measure can be defined based on their approach with important gains compared to [8],[9] in particular and other state-of-the-art QMs [3],[11] in general. Relevant aspects of their work are presented in the following section, while the proposed improvements will be described in Sec. 2.3.

2.2 Preliminary Considerations

The Video Quality Measure (VQM) proposed in [8],[9] is a top-down approach that integrates some salient properties of the HVS. This type of approach is preferred in this work in order to avoid detailed formulation of assumptions on sparsely understood functional properties of early vision stages.

The mathematical formulation of the VQM for single pictures, $Q(t)$, is given as

1-4244-1437-7/07/$20.00 ©2007 IEEE

$$\mathrm{Q}(t) = \frac{\mathrm{e}^{\frac{\gamma}{1+\delta(t)}}}{\mathrm{e}^{\gamma}} \tag{1}$$

where the term γ can be freely selected and steers the interval of $\delta(t)$ for which the contrast is enhanced or reduced. The denominator in (1) is a normalization factor. The variable $\delta(t)$ represents a differential term that assesses the distance between a given reference and a corresponding distorted signal. The definition of $\delta(t)$ is given in (2), where $\mathrm{E}_\mathrm{o}(t)$ and $\mathrm{E}_\mathrm{d}(t)$ correspond to mean absolute spatial gradients in the original and the distorted signal respectively. $\delta(t)$ typically features values that lie in the interval [0,1]. However, the latter can not be guaranteed as slight overshoots may occur depending on the data. Given the range of $\delta(t)$, the range of $\mathrm{Q}(t)$ can be given as $[\mathrm{e}^{-0.5\gamma},1]$. Note that $\mathrm{Q}(t)$ must be maximized to ensure good video quality at the decoder side.

$$\delta(t) = \frac{\left|\mathrm{E}_\mathrm{o}(t) - \mathrm{E}_\mathrm{d}(t)\right|}{\mathrm{E}_\mathrm{o}(t)} \tag{2}$$

$\mathrm{E}_\mathrm{o}(t)$ and $\mathrm{E}_\mathrm{d}(t)$ are defined as

$$\mathrm{E}_\mathrm{o}(t) = \frac{1}{M(\frac{N}{4}-1)} \sum_{\xi=1}^{M} \sum_{\eta=1}^{\frac{N}{4}-1} \left|\mathrm{o}_{0^\circ}(\xi,4\eta,t)\right| + \frac{1}{(\frac{M}{4}-1)N} \sum_{\xi=1}^{\frac{M}{4}-1} \sum_{\eta=1}^{N} \left|\mathrm{o}_{90^\circ}(4\xi,\eta,t)\right| \tag{3}$$

$$\mathrm{E}_\mathrm{d}(t) = \frac{1}{M(\frac{N}{4}-1)} \sum_{\xi=1}^{M} \sum_{\eta=1}^{\frac{N}{4}-1} \left|\mathrm{d}_{0^\circ}(\xi,4\eta,t)\right| + \frac{1}{(\frac{M}{4}-1)N} \sum_{\xi=1}^{\frac{M}{4}-1} \sum_{\eta=1}^{N} \left|\mathrm{d}_{90^\circ}(4\xi,\eta,t)\right| \tag{4}$$

$$\mathrm{d}_\beta(\xi,\eta,t) = \left[\mathrm{d}(\xi,\eta,t) * \mathrm{f}_\beta(\xi,\eta)\right]\mathbf{C}(\xi,\eta,t) \;\wedge\; \mathrm{o}_\beta(\xi,\eta,t) = \left[\mathrm{o}(\xi,\eta,t) * \mathrm{f}_\beta(\xi,\eta)\right]\mathbf{C}(\xi,\eta,t) \tag{5}$$

Eqs. (3) and (4) reflect the mean occurrence of given gradient directions in the original signal $\mathrm{o}(\xi,\eta,t)$ and the distorted signal $\mathrm{d}(\xi,\eta,t)$. $\mathrm{o}_\beta(\xi,\eta,t)$ and $\mathrm{d}_\beta(\xi,\eta,t)$ represent the high-pass filtered (orientation $\beta = (0°, 90°)$) original and distorted signals respectively. (M,N) depict the width and the height of the video signal. The filtering operation is defined in (5), where $\mathrm{f}_\beta(\xi,\eta)$ is a linear, anisotropic gradient filter of orientation β used for edge detection. * is the convolution operation. $\mathbf{C}(\xi,\eta,t)$ represents the object contour matrix that is determined from the original signal. This matrix emphasizes regions of large spatial contrast as object boundaries in the edge masks, as the former are assumed to be of salient relevance for subjective quality perception [8],[9]. $\mathbf{C}(\xi,\eta,t)$ is determined as an isotropic edge mask. The single picture VQM formulation in (1)-(5) corresponds to the block fidelity (BF) measure proposed by Ong et al. in [8],[9]. A simple gradient filter is used in their proposal. Their VQM is furthermore constrained to the block-based video coding framework. 4x4 macroblock transitions are considered in this formulation in order to detect possible blocking effects (3),(4).

Ong et al. [8],[9] sample the high-pass masks in parallel to the gradient direction, i.e. horizontal edge masks are sampled vertically while vertical edge masks are sampled horizontally. Additionally to BF, their approach comprises two further measures that aim to capture properties of the HVS (e.g. masking). The latter measures are pooled with BF in a multiplicative manner. One of the additional measures, the so-called distortion invisibility term, however, is very complex as it incorporates color, spatial and temporal masking. The formulation of the different masking properties features more than 20 degrees of freedom, which appears to be hardly manageable.

2.3 Proposed Improvements for Block Transform VC

2.3.1. Gradient Filter

The first modification proposed in this work applies to the gradient filter used in (5). A filter like Sobel is suggested to achieve robustness against spurious edges. The selected edge detector approximates the gradient of local luminance. Sobel has smoothing abilities, which makes it relatively insensitive to noise [7].

2.3.2. Feature Point Selection

Disturbing artifacts like blocking can be detected by the block fidelity measure as already explained above. As the contrast of the pictures plays an important role in quality perception [6], it is suggested to build this into the block fidelity measure. It should be noted that the block fidelity measure formulated by Ong et al. [8],[9] inherently detects blurring artifacts, when high frequency areas that are particularly affected by low-pass effects are sampled.

The case that $\mathrm{E}_\mathrm{d}(t)$ is larger than $\mathrm{E}_\mathrm{o}(t)$ in (2) indicates that the distorted signal has gained some additional edges compared to the original picture. This may be related to tiling effects in the distorted signal. When $\mathrm{E}_\mathrm{d}(t)$ is smaller than $\mathrm{E}_\mathrm{o}(t)$, it can be assumed that a loss of contrast has occurred between original and distorted signal. The contrast loss detection property of the block fidelity measure is however limited by the fact that structural features like object boundaries are not properly sampled in (3),(4).

Let Fig. 1a exemplarily depict an artificial image that is to be analyzed using the block fidelity measure. Fig. 1b then depicts the response obtained by applying the sampling approach described in (3),(4). The sampled images are the basis of the block fidelity measure. It can be seen that major structural information is ignored by this measure if the former does not match the macroblock grid, which will assumingly be the case in most natural images. In this work, a sampling of the edge masks orthogonally to the gradient's direction is proposed, which yields better feature points (cf. Fig. 1c). Object boundaries that are particularly important for subjective quality perception [8],[9] are better preserved by the new approach. The blocking detection feature points depicted in Fig. 2 (bottom, gray samples) constitute a subset of the overall set of feature points selected for quality assessment in this scenario. Notice that object boundaries are represented by the bright samples in Fig. 2 (bottom, white samples).

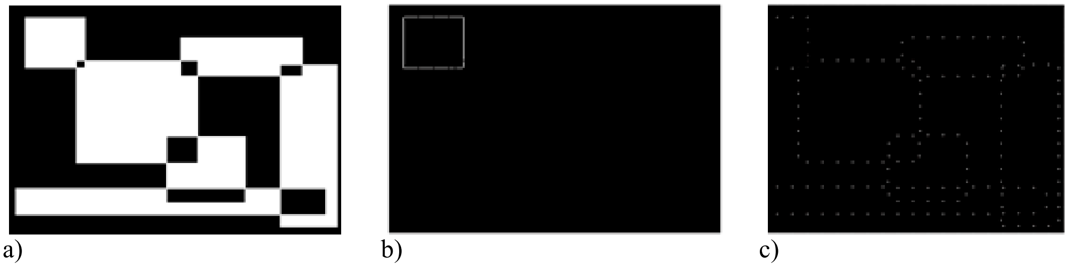

Fig. 1 – Improved feature point selection for image quality assessment, a) Example of an input image to be analyzed, b) Inaccurate sampling of structural information by Ong et al. [8],[9], c) Accurate sampling of structural information by the proposed VQM

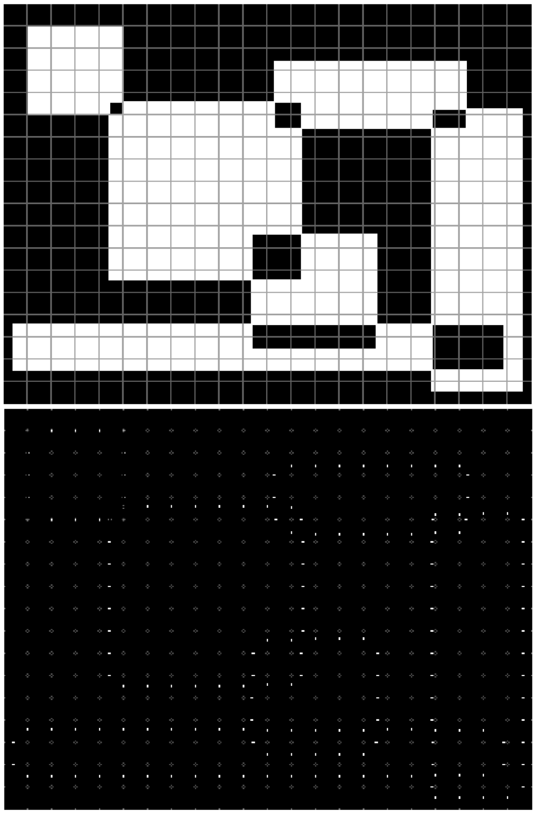

Fig. 2 – Accurate sampling of structural information and macroblock transitions by the proposed VQM. Top: Input image to be analyzed, bottom: Sampled input picture

The proposed feature point selection can be formalized as follows

$$\mathrm{E_o}(t)=\frac{1}{(\frac{M}{\kappa}-1)N}\sum_{\xi=1}^{\frac{M}{\kappa}-1}\sum_{\eta=1}^{N}\left|\mathrm{o}_{0^\circ}(\kappa\xi,\eta,t)\right|\mathrm{m}(\xi,\eta,t)+\frac{1}{M(\frac{N}{\kappa}-1)}\sum_{\xi=1}^{M}\sum_{\eta=1}^{\frac{N}{\kappa}-1}\left|\mathrm{o}_{90^\circ}(\xi,\kappa\eta,t)\right|\mathrm{m}(\xi,\eta,t) \quad (6)$$

$$\mathrm{E_d}(t)=\frac{1}{(\frac{M}{\kappa}-1)N}\sum_{\xi=1}^{\frac{M}{\kappa}-1}\sum_{\eta=1}^{N}\left|\mathrm{d}_{0^\circ}(\kappa\xi,\eta,t)\right|\mathrm{m}(\xi,\eta,t)+\frac{1}{M(\frac{N}{\kappa}-1)}\sum_{\xi=1}^{M}\sum_{\eta=1}^{\frac{N}{\kappa}-1}\left|\mathrm{d}_{90^\circ}(\xi,\kappa\eta,t)\right|\mathrm{m}(\xi,\eta,t) \,. \quad (7)$$

where a $\kappa \mathrm{x} \kappa$ macroblock size is assumed and $\mathrm{m}(\xi,\eta,t)$ is a binary mask that defines the regions of interest. Its components are set to one if such a region is given and set to zero otherwise. Notice that due to the purposeful inclusion of boundary samples into the set of feature points used for quality assessment, not only can tiling effects be detected by the new measure but also distortions affecting object boundaries. Hence, the initial block fidelity measure has been generalized to achieve a simple, generic impairment detection tool. The global quality score for an entire video sequence is determined as the median of the single picture qualities.

3. EXPERIMENTAL RESULTS

Experimental evaluations are conducted to validate the proposed video quality measure for block-based hybrid video coding. Validation is done with corresponding ground truth sets. The proposed spatial VQM is thereby compared to relevant impairment measures in order to benchmark its performance. All quality measures are evaluated by matching the corresponding predicted opinion scores with subjective Differential Mean Opinion Scores (DMOS).

3.1 Test Framework

The proposed model is evaluated based on two major criteria recommended by VQEG [11]: Pearson's (r_p, prediction accuracy) and Spearman's correlation coefficients (r_s, performance w.r.t. the relative magnitudes of subjective quality ratings). In order to ensure reproducibility of the results achieved in the experiments, the cross validation approach [10] is used to determine r_p and r_s.

Unfortunately, the official VQEG (Phase II) test data [11] were not accessible to the authors. Hence, two alternative ground truths have been used in the present work to evaluate the proposed quality measure. The data sets are provided by MPEG [12] and Fraunhofer Heinrich-Hertz-Institut (HHI). The MPEG data set consists of four video sequences (QCIF, 10Hz and 15Hz, 10s, 32 kbps and 64kbps) and was formerly used to benchmark the performance of MPEG-4 and H.26L anchors [12]. The HHI data correspond to five video clips (QVGA, 12.5 Hz, 10s, variable bit rate) obtained from several German television channels. The clips feature various contents as news, sports, cartoon, monochrome and colour movies that are MPEG-2 coded.

3.2 Results

In this section, the proposed video quality measure is evaluated w.r.t. its performance and compared to PSNR and

VQEG's best quality assessor proposed by the National Telecommunications and Information Administration (NTIA) [5]. The full image plane is considered for quality assessment. The results are summarized in Fig. 3 and Fig. 4. As can be seen, the proposed VQM yields similar r_s and r_p results to NTIA's VQM. For the sake of completeness, the results for the measure proposed by Ong et al. [8],[9] (cf. Sec. 2.2) are also given. It can be seen that the latter VQM performs significantly worse than NTIA's and the proposed quality assessors. It should be noted that Ong et al.'s VQM features more than 20 degrees of freedom. It was not possible to obtain the optimized parameter settings due to patent issues. Hence, default parameters were set to the best of our knowledge in the present work ($f_{BF}=f_{BF}=0.25$, $\gamma_1=\gamma_2=10^{-10}$, $s=T_{m,2}=L_m=\alpha_4=T_{l,2}=L_l=0$, $C^{is}=f_s=T_{m,o}=T_{m,1}=\alpha_1=\alpha_2=\alpha_3=T_{l,o}=T_{l,1}=1$, $T_{m,3}=r=0.1$, $z_1=z_2=\nu_1=\nu_2=e$).

The per picture complexities of the relevant VQMs are estimated via a Taylor series expansion

$$\begin{aligned} &Pr_{PSNR} = 2MN + 6, Add_{PSNR} = 3MN \\ &Pr_{NTIA} = 169\,MN, Add_{NTIA} = 168\,MN \\ &Pr_{VQM} \approx 5.87MN + 3, Add_{VQM} \approx 2.625MN + 2 \end{aligned} \tag{8}$$

where Pr and Add represent the number of products/divisions and additions/subtractions, respectively.

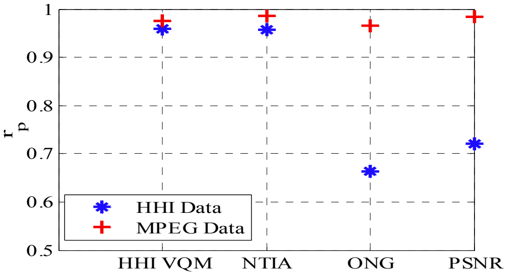

Fig. 3 – Pearson's correlation coefficients for evaluated VQMs

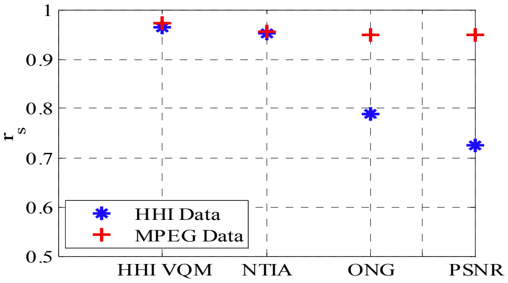

Fig. 4 –Spearman's correlation coefficients for evaluated VQMs

A description of NTIA's general VQM can be found in [5]. It corresponds to a linear combination of 7 terms of which only the most complex (si_loss) is considered here. Hence the estimation above represents a lower bound. It can be seen in (8) that the proposed quality assessor is significantly less complex than NTIA's general VQM and slightly more complex than PSNR.

4. CONCLUSIONS AND FUTURE WORK

The quality assessment tool presented in this paper provides the performance of VQEG's best quality measure (NTIA) at roughly twice the complexity of PSNR, which shows much lower performance. In fact, the proposed VQM always yields a slightly higher correlation coefficient (Spearman) than VQEG's NTIA measure. The proposed measure can be easily extended to other applications than the coding scenario. This is achieved by adapting the definition of the mask $m(\xi,\eta,t)$ (cf. (6),(7)) and the feature point selection to the requirements of the given application. Some HVS properties have not been considered in the present work. Hence, a better representation of, for instance, multi-resolution properties of the HVS [4] may help improve the performance of the proposed video quality measure.

5. REFERENCES

[1] TU-R WG6Q 6/39-E, "Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference", October 2003.

[2] A B. Watson and J. Malo, "Video Quality Measures Based on the Standard Spatial Observer", *Proc. ICIP 2002*, Vol. 3, p. 41-44, Rochester, New York, USA, 2002.

[3] B. Girod, "What's Wrong with Mean-Squared Error", *Digital Images and Human Vision*, Ed. Cambridge, MA: MIT Press, p. 207-220, 1993.

[4] S. Winkler, "Issues in Vision Modeling for Perceptual Video Quality Assessment", *Elsevier Signal Processing*, Vol. 78, p. 231-252, 1999.

[5] M. H. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality", *IEEE Transactions on Broadcasting*, Vol. 50, No. 3, p. 312-322, September 2004.

[6] Z. Wang, A. Conrad 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, April 2004.

[7] J.-R. Ohm, "Multimedia Communication Technology", ISBN 3-540-01249-4, *Springer*, Berlin Heidelberg New York, 2004.

[8] E. P. Ong, W. Lin, Z. Lu, and S. Yao, "Colour Perceptual Video Quality Metric", *Proc. ICIP 2005*, Vol. 3, p. 1172-1175, Genova, Italy, 2005.

[9] E. P. Ong, X. Yang, W. Lin, Z. Lu, and S. Yao, "Video Quality Metric for Low Bitrate Compressed Video", *Proc. ICIP 2004*, p. 3531-3534, Singapore 2004.

[10] C. M. Bishop, "Neural Networks for Pattern Recognition", ISBN 0198538642, *Oxford University Press*, 1995.

[11] ITU-R WG6Q 6/39-E, "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II", August 2003.

[12] ISO/IEC JTC 1/SC 29/WG 11 N3671, "Call for Proposals for New Tools to further Improve Video Coding Efficiency", La Baule, France, October 2000.