BLIND QUALITY ASSESSMENT FOR JPEG2000 COMPRESSED IMAGES

Hamid R. Sheikh, Zhou Wang, Lawrence Cormack and Alan C. Bovik

Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering,
The University of Texas at Austin, Austin, TX 78712-1084, USA.
Email: {hamid.sheikh, zhouwang}@ieee.org, cormack@psy.utexas.edu, bovik@ece.utexas.edu

ABSTRACT

Measurement of image quality is crucial for many image-processing algorithms, such as acquisition, compression, restoration, enhancement and reproduction. Traditionally, image quality assessment algorithms have focused on measuring image fidelity, where quality is measured as fidelity with respect to a 'reference' or 'perfect' image. The field of blind quality assessment has been largely unexplored. In this paper we present an algorithm for blindly determining the quality of JPEG2000 compressed images. Our algorithm assigns quality scores that are in good agreement with data from human observers. Our algorithm utilizes a statistical model for wavelet coefficients and computes features that exploit the fact that quantization produces more zero coefficients than expected for natural images. The algorithm is trained and tested on data obtained from human observers, and performs close to the limit on useful prediction imposed by the variability between human subjects.

1. INTRODUCTION

With the advent of digital content, the problem of automatically quantifying its quality has received tremendous attention in the research community. Digital images are now a part of our everyday lives, and the need to discover ways of assessing their quality in a way that is consistent with human assessment has led to several approaches towards solving the problem. One class of quality assessment schemes is the class of image fidelity metrics, which assume that a 'reference' image is available against which to compare a distorted or processed image against. However, human observers can readily judge the quality of images without explicit reference images. We were thus inspired to consider *blind quality assessment*, in which an algorithm seeks to assign quality scores that are consistent with human perception but without an explicit comparison with the reference image.

Blind Quality Assessment is a very hard problem since many unquantifiable factors play a role in human assessment of quality, such as aesthetics, cognitive relevance, learning, context etc. These factors introduce variability among human observers in judgements of image quality. However, we can work with the following philosophy for blind image quality assessment: *all images are perfect, regardless of content, until distorted by acquisition, processing or reproduction.* Hence, the task of blind quality measurement simplifies into measuring the distortion that has possibly been introduced in the image during the stages of acquisition, processing or reproduction. The reference for measuring this distortion would be the statistics of 'perfect' natural images, measured with respect to a model that best suits a given distortion type or application. This philosophy effectively decouples the unquantifiable aspects of image quality mentioned above from the task of objective quality assessment. All 'perfect images' are treated equally, disregarding the amount of cognitive information in the image or its aesthetic value.

One class of image processing systems that introduces distortions in images is the class of the lossy image compression algorithms, which reduce the storage/transmission bandwidth requirements (beyond the limits of lossless compression) by throwing away information that is of least visual relevance [1]. However, once the compression ratios increase beyond a certain limit, the distortions become perceptible, and even annoying. One very popular image compression scheme is JPEG, which introduces a very distinct compression artifact, the blocking artifact, which occurs because JPEG processes images as non-overlapping 8×8 blocks with the Discrete Cosine Transform (DCT), and hence introduces discontinuities in the image at block boundaries. Researchers have previously reported their success at measuring the blocking artifact blindly, and thereby quantifying the image quality for JPEG compressed images (or DCT based compressed video) without the reference signals [2, 3]. In [4], a blocking measure is combined with

H. R. Sheikh and A. C. Bovik are affiliated with the Laboratory for Image and Video Engineering, and L. Cormack is affiliated with the Center for Perceptual Systems, The University of Texas at Austin. Z. Wang is affiliated with The Laboratory for Computational Vision, New York University, New York.

This research was supported in part by Texas Instruments, Inc., and by State of Texas Advanced Technology Program.

estimates of image activity to give quality scores to images.

JPEG2000 is a recent standard that performs better than the JPEG algorithm by providing higher compression ratios at similar visual quality. JPEG2000 uses the Discrete Wavelet Transform instead of the DCT, but at higher compression ratios, it too introduces distortions that typically include blurring and ringing. While some previous work has reported attempts to measure the ringing artifact in Wavelet based image coders [5], the ringing artifact metric has not been calibrated or tested against human judgements. Also, our observations demonstrate that the ringing artifacts appear only for a narrow range of compression ratios, and in lightly textured regions around strong edges. Blurring is the more dominant artifact for JPEG2000 (and wavelet based coders), especially at lower bit rates. We believe that the work presented in this paper is the first attempt to blindly assess the quality of images compressed by JPEG2000 (or any other wavelet based) image compression systems in a way that is directly related to human perception of quality, with our algorithm calibrated and tested against human subject data.

In our research, we make use of a statistical model for natural images in the wavelet domain to quantify the loss in quality due to quantization of wavelet coefficients. Using this model, we extract features that capture the quantization process in the wavelet domain. Using quality judgements of 198 images (29 uncompressed images, 169 compressed images) by 25 human subjects, we train our model to assign scores to images that directly relate to human perceptions.

2. SUBJECTIVE EXPERIMENTS

Twenty-nine high-resolution 24-bits/pixel RGB color images (typically 768×512) were compressed using JPEG2000 with different compression ratios to yield a database of 198 images, 29 of which were the original (uncompressed) images. The bit rates used for compression were in the range of 0.03 to 3.2 bits per pixel, chosen such that the resulting distribution of quality scores was roughly uniform over the entire range. Observers were asked to provide their perception of quality on a continuous linear scale that was divided into five equal regions marked with adjectives "Bad", "Poor", "Fair", "Good" and "Excellent". The testing was done in two sessions with about 25 subjects per session evaluating about half of the images. Hence, about 25 human observers rated each image. The raw scores for each subject were normalized by the mean and variance of that subject (that is, converted to Z-scores [6]) and then scaled and shifted by the mean and variance of the entire subject pool to the full range (1 to 100). Mean scores were then computed for each image after removing outliers. The average standard deviation of the scores for all images with the mean score used for training was found to be 6.8 (on a scale of 1-100). The average value of the linear correlation coefficient of subjects with the mean score of 0.93.

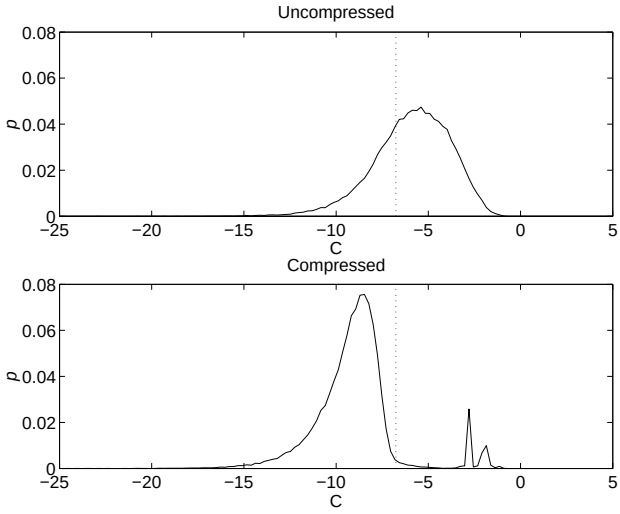

Fig. 1. Histograms of the logarithm (to the base 2) of horizontal wavelet coefficient at the finest scale for one image, before and after compression (at 0.75 bits per pixel). The dotted line denotes the threshold at -6.76.

3. STATISTICAL MODEL FOR IMAGES IN THE WAVELET DOMAIN

We have used a statistical model for natural images proposed in [7, 8]. It captures the statistics of wavelet coefficients in a given subband and their correlations with other wavelet coefficients in different subbands. We observed from our experiments that this model is suitable for measuring the effect of quantization of wavelet coefficients, since quantization pushes wavelet coefficients at finer scales towards zero. This results in a greater probability of zero coefficients in any subband than expected for natural images.

The statistical model proposed in [7, 8] for the probability density function of the wavelet coefficient's magnitude, C, conditioned on the magnitude of the linear prediction of the coefficient, P, is given in (1) where M and N are assumed to be independent zero mean random variables.

$$C = MP + N \tag{1}$$

[7, 8] use an empirical distribution for M and assumes N to be Gaussian of unknown variance. In our method, as a first approximation we consider the marginal distribution of the wavelet coefficients only. Figure 1 shows the histogram of the logarithm (to the base 2) of the horizontal wavelet coefficient magnitude for one image, and the histogram for the same compressed image at the same scale and orientation. Quantization shifts the histogram towards lower values.

We divide the histogram into two regions: insignificant coefficients and significant coefficients. We found that the probability of a coefficient being in one of the regions at a certain scale and orientation is a good feature to represent the effect of quantization. The thresholds for the coefficients are determined empirically, such that the error in quality prediction is minimized over the training set.

4. FEATURES FOR QUALITY ASSESSMENT

We only work with the luminance component of the images in our algorithm, which was normalized to a constant root-mean-squared value of 1.0. Let $h(C)$ denote the normalized histogram of the logarithm of the magnitude of wavelet coefficients at a given scale and orientation, where the edge bins extend up to infinity in both directions. We estimate the probability that the coefficient magnitude is significant (above the threshold T_h), denoted by p_s, from the normalized histogram: $p_s = \sum_{C>T_h} h(C)$

The feature vector that we chose for our algorithm is $\{p_{si}\}$ where i is an index into the wavelet subband for which the probability is being calculated. In our experiment, we chose the detail coefficients at the two finest scales for the horizontal, vertical and diagonal orientations. We reduce the dimensionality of the six-dimensional vector by Principal Component Analysis (PCA) to just one component, p_w, and use the fit in (2) to relate the feature to the quality predictions, Q_P, where c_i denotes the elements of the first PCA basis vector, and μ_{si} denotes the elements of the mean vector.

$$\begin{aligned} p_w &= \sum_{i=1}^{6} c_i \left(p_{si} - \mu_{si}\right) \\ Q_P\left(p_w\right) &= K\left(1 - \exp\left(-\frac{(p_w - u)}{T}\right)\right) \end{aligned} \quad (2)$$

The fitting equation is a saturating exponential, where u is a shift parameter, T denotes the decay constant, K denotes the highest quality score that the algorithm can possibly give.

5. TRAINING AND TESTING

The database of 29 original images (and their corresponding compressed versions) was split randomly into two groups: 15 training images (and their compressed versions) and 14 test images (and their compressed versions). The biorthogonal $9/7$ wavelet [8] was used for the decomposition. The parameters computed during training are the thresholds, the PCA basis vectors and the fitting parameters for Q_P. In our training, the thresholds and the best-fit parameters were obtained using multidimensional unconstrained nonlinear minimization (MATLAB command *fminsearch*) to minimize the prediction error over the training set.

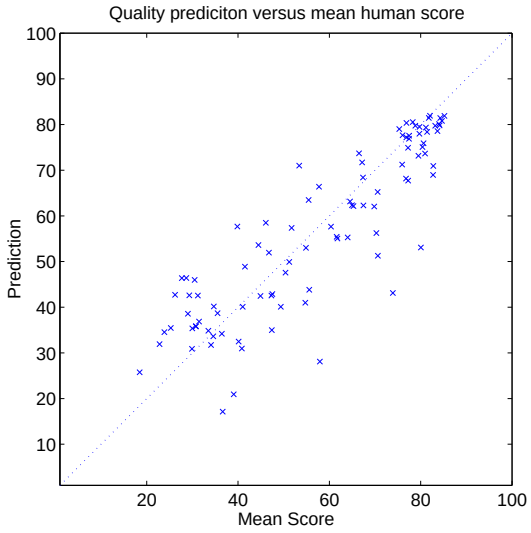

Fig. 2. Quality predictions versus mean human score.

The quality predictions on the test data set were evaluated using the parameters learned from the training set. These predictions were compared against the actual human evaluations. Figure 2 shows the correlations between these predictions and the human ratings for one run on the test data. Figure 3 shows the histogram of the root-mean-squared-error (RMSE) between the predictions and the mean human scores for the testing set for several runs of the algorithm (each time with a different, and random, training and test subsets of the database). The RMSE values for the prediction errors should be compared with the average standard deviation of 6.8 for the human scores versus the mean score for all images. This shows that our algorithm performs close to the limit on useful prediction imposed by the variability within human observers. Running multiple tests with different training and test configurations shows that the algorithm's performance is stable.

Table 1 gives the values of the mean thresholds from several runs, Table 2 gives the mean PCA vectors and Table 3 gives the mean values of the parameters for the fit. Here, i goes from 1 to six to index the following subbands: horizontal, vertical and diagonal details at the second finest scale, horizontal, vertical and diagonal details at the finest scale.

6. CONCLUSIONS

In this paper we have presented an algorithm for blindly determining the quality of images that have been compressed by JPEG2000. As far as we are aware, this is the first attempt of its kind to design an algorithm for blindly evaluating the quality for JPEG2000 compressed images. The al-

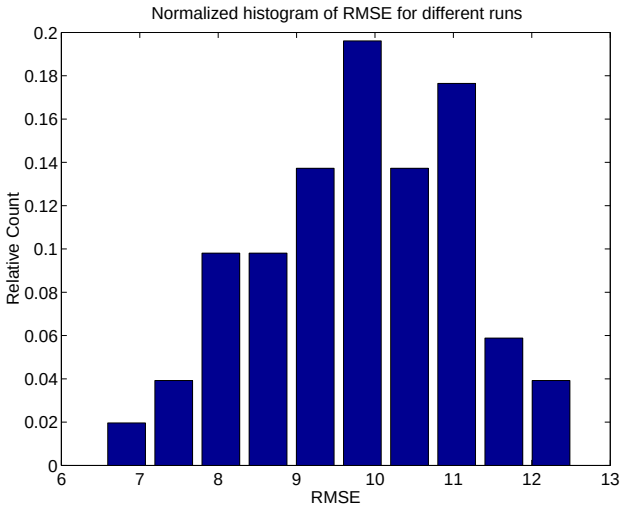

Fig. 3. RMSE values for the test data for different runs.

Subband	Threshold
H2	-6.354 ± 0.673
V2	-6.300 ± 0.921
D2	-6.250 ± 0.605
H1	-6.049 ± 1.988
V1	-4.927 ± 0.717
D1	-4.928 ± 0.520

Table 1. 1D simplification: Mean thresholds for C in the log_2 domain for several runs of the algorithm, with $\pm 1\sigma$. The thresholds are shown for the horizontal, vertical and diagonal orientations at the 2^{nd} finest and finest resolutions.

i	μ_{si}	c_i
1	0.266 ± 0.027	0.452 ± 0.048
2	0.233 ± 0.028	0.425 ± 0.061
3	0.285 ± 0.215	0.372 ± 0.081
4	0.174 ± 0.050	0.442 ± 0.043
5	0.168 ± 0.070	0.403 ± 0.051
6	0.096 ± 0.036	0.313 ± 0.061

Table 2. Principle Component vectors computed from the training data for several runs of the algorithm, with $\pm 1\sigma$.

Parameter	
K	82.236 ± 0.926
u	-0.584 ± 0.041
T	0.323 ± 0.040

Table 3. Parameters for computing output quality from (2)

gorithm utilizes a statistical model for wavelet coefficients and computes features that exploit the fact that quantization of wavelet coefficients produces more zero coefficients than expected in natural images. The probabilities of the coefficients being non-zero in different subbands is used as a feature, together with PCA based dimensionality reduction and non-linear curve-fitting to do predictions of image quality scores. The algorithm is trained and tested on data obtained from human observers. On a scale of 1-100, an average RMSE of approximately 9.8 between quality predictions and human evaluations is reported, which is close to the average standard deviation of 6.8 for quality scores assigned by human observers. More research needs to be conducted to reduce this gap. We are continuing research into using higher-order models of natural image statistics in the wavelet domain to achieve this goal.

7. REFERENCES

[1] A. C. Bovik, *Handbook of Image and Video Processing*. Academic Press, 2000.

[2] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters*, vol. 4, pp. 317–320, Nov. 1997.

[3] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Proc.*, (Vancouver, Canada), pp. 981–984, Oct. 2000.

[4] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. Image Proc.*, Sept. 2002.

[5] S. Yang, Y. H. Hu, T. Q. Nguyen, and D. L. Tull, "Maximum-likelihood parameter estimation for image ringing-artifact removal," *IEEE Trans. on Circuits and Systems for Video Tech.*, vol. 11, pp. 963–973, Aug. 2001.

[6] A. M. van Dijk, J. B. Martens, and A. B. Watson, "Quality assessment of coded images using numerical category scaling," *Proc. SPIE*, vol. 2451, pp. 90–101, Mar. 1995.

[7] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, Nov. 1997.

[8] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. on Image Processing*, vol. 8, pp. 1688–1701, Dec. 1999.