ADAPTIVE SCALES AS A STRUCTURAL SIMILARITY INDICATOR FOR IMAGE QUALITY ASSESSMENT

Atanas Boev, Alessandro Foi, Karen Egiazarian, Vladimir Katkovnik

Tampere University of Technology, Tampere, Finland

ABSTRACT

Existing image quality metrics, based on the concept of "structural similarity" (e.g. [1][2]) fail on particular image impairments. We propose the use of "similarity of adaptive scales" as the key indicator of structural similarity, thus merging scale and structure in one concept. Such adaptive scales can be effectively determined by a modification the intersection of confidence intervals (ICI) algorithm [6], originally developed for image restoration. We show that changes in the image structure are reflected into changes in the adaptive scales, and that the adaptive scales are more sensitive to structural changes than the conventional measures based on local statistics.

We present some simulation experiments demonstrating the improvement over the current structural similarity measures, and propose a novel, adaptive-scale quality metric. This metric is obtained by combining a term based on adaptive-scales similarity with another term taking into account differences in the image intensity. This metric has been tested on large databases [7] with subjective mean-opinion scores for JPEG, JPEG2K, blur, and white noise degradations, showing a close agreement of the proposed metric with the quality as perceived by human evaluators.

1. INTRODUCTION

For a long time, there have been two approaches for quality estimation – *objective*, where evaluations are based on a mathematical model, and *subjective*, where the quality is judged by a group of human observers. In the case of *full-reference* image quality assessment, the quality is expressed as a measure of the similarity between the "original" image (assumed to have perfect quality) and the test image. In order to have correlation with the subjective opinion, the similarity measure should be relevant to the processes in the human visual system.

Structural features of the image are dominant in traditional quality metrics [4][5]. The contemporary concept of using "structural similarity" as quality indicator is based on the assumption that human vision is highly optimized for extracting structural information [1].

However, the current state-of-the-art metrics following the concept (e.g. SSIM in [1] and [2]) produce unsatisfactory results on particular image impairments – for example blur or sharpening. We identify the lack of proper scale-adaptivity as one of the reasons for such inadequacy. Indeed, modern models of the human perception assert that vision is intrinsically multiscale [3], which suggest that a "similarity of scales" would be an important element for full-reference perceptual quality evaluation. However, the use of the multiscale approach has been rather limited. Recent works [2] aimed mainly at improving the flexibility of existing monoscale "structural similarity" metrics with respect to different viewing conditions, rather than directly exploiting the scales as a component of the quality metric.

In Figure 1 (a-d), we present image artifacts which have the same mean SSIM quality index and obviously different visual appearance.

2. ADAPTIVE SCALES AS STRUCTURAL SIMILARITY INDICATOR

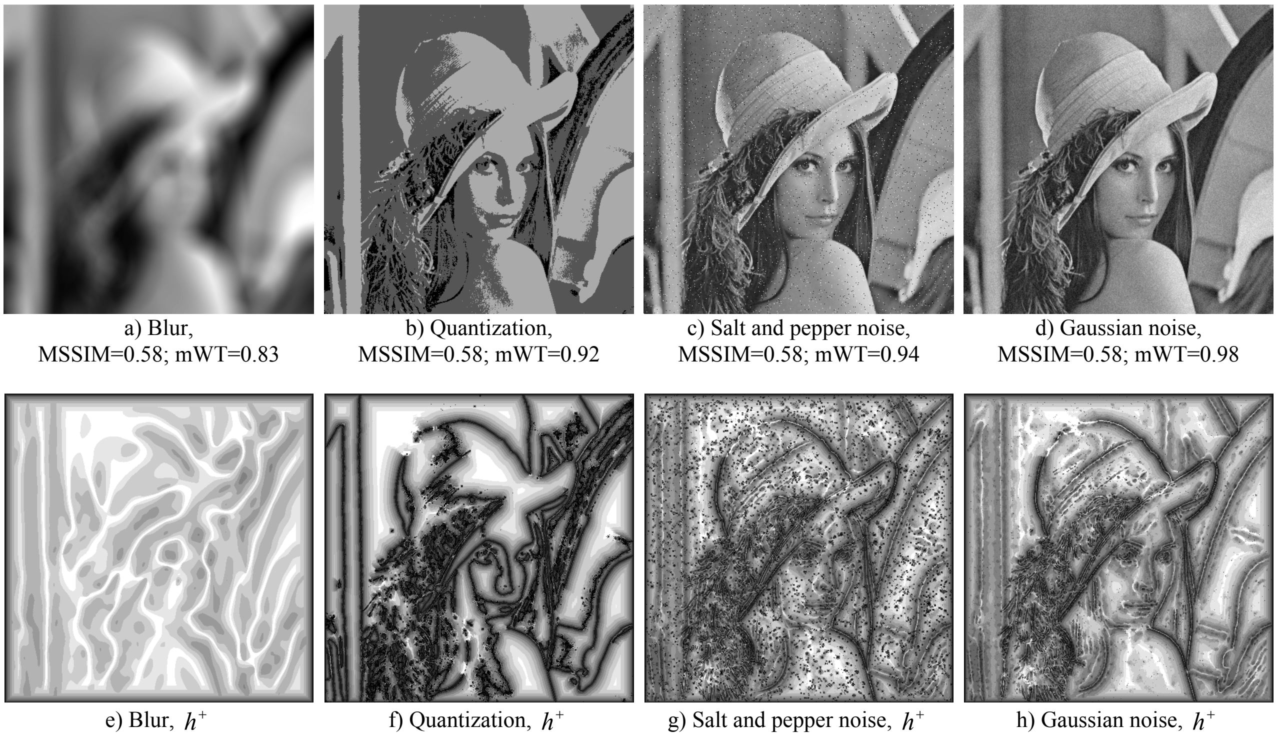
We consider a generalized version of the intersection of confidence intervals (ICI) rule [6] for adaptive-scale selection.

Let $\{g_h\}_{h\in H}$ be a collection of varying-scale smoothing kernels $g_h(\cdot)=g(\cdot/h)/h^2$, where the parameter $h>0$, known as *scale or bandwidth* of g_h, controls the smoothing effect of the kernels.

Given a signal or an image y, the kernels $\{g_h\}_{h\in H}$ are used as convolutional filters yielding a collection $\{\hat{y}_h\}_{h\in H}$, $\hat{y}_h = y * g_h$, of differently smoothed versions of y. Usually, the kernel g is a positive monomodal symmetric function, centered in the origin. It is always normalized such that $\int g = 1$, ensuring preservation of the mean.

The smoothing effect of the kernels can be quantified explicitly by the kernel's l_2-norm. For the particular case of Gaussian noise degradation, the l_2-norm of the kernel is exactly the amplification factor for the standard deviation of the noise after convolutional filtering. It is easy to verify that $\|g_h\|_2$ decreases as h increases, thus a larger scale h corresponds to a stronger smoothing.

Given a finite ordered set of scales $H=\{h_1<\cdots<h_j\}$ and the corresponding $\{\hat{y}_{h_j}(x)\}_{j=1}^{J}$ we determine a sequence $\{D_j\}_{j=1}^{J}$ of "confidence" intervals



a) Blur, MSSIM=0.58; mWT=0.83
b) Quantization, MSSIM=0.58; mWT=0.92
c) Salt and pepper noise, MSSIM=0.58; mWT=0.94
d) Gaussian noise, MSSIM=0.58; mWT=0.98

e) Blur, h^+
f) Quantization, h^+
g) Salt and pepper noise, h^+
h) Gaussian noise, h^+

Fig. 1. Mean structural similarity index (MSSIM) and ICI Window Term (WT) of four different impairments (a-d); the corresponding adaptive scales (e-h).

$$D_j = \left[\ \hat{y}_{h_j}(x) - \Gamma\left\|g_{h_j}\right\|_2, \hat{y}_{h_j}(x) + \Gamma\left\|g_{h_j}\right\|_2\ \right], \quad (1)$$

where $\Gamma > 0$ is a fixed threshold parameter and x denotes a fixed pixel coordinate in the image domain X. The modified ICI rule for adaptive-scale selection can be stated as follows:

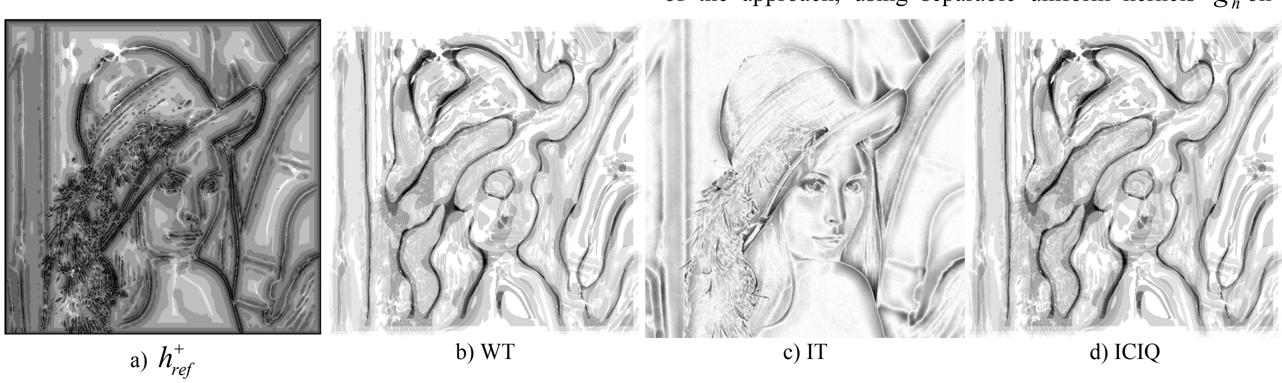
Consider the intersection of intervals $I_j = \bigcap_{i=1}^{j} D_i$ and let j^+ be the largest of the indexes j for which I_j is non-empty, $I_{j^+} \neq \varnothing$ and $I_{j^+ +1} = \varnothing$. The adaptive scale h^+ is defined as $h^+ = h_{j^+}$.

This is a procedure for a fixed x. It is produced for all $x \in X$ and in this way we obtain a pointwise-adaptive scale $h^+ = h_{j^+}$ for every pixel in the image.

Roughly speaking, the ICI selects the coarsest scale estimate which is compatible with all finer scales. In practice this means that adaptively, for every pixel, the criterion allows the maximum degree of smoothing, stopping before oversmoothing begins.

Let us remark that this generalized ICI concides with the usual ICI [6][9], as it is defined for the particular case of observations with Gaussian noise, provided that the standard deviation of the noise is incorporated into the threshold parameter Γ.

In what follows we consider the most basic implementation of the approach, using separable uniform kernels g_h on a



a) h^+_{ref}
b) WT
c) IT
d) ICIQ

Fig. 2. Building elements of the ICIQ metric: a) Adaptive scales for the reference image, b) Window Term for the test image, c) Intensity Term for the test image, d) ICIQ. (The test image is shown in Figure 1a and its adaptive scales are shown in Figure 1e)

square support of size $h \times h$. This allows for a fast computation of the smoothed $\{\hat{y}_h\}_{h \in H}$ through a recursive filterbank structure.

Although the adaptive scales h^+, do not portray any explicit information about the actual image intensities, they contain significant knowledge about the structures in the image, accurately revealing edges, contours and intensity gradients, as shown in the examples in Figures 1e-1h. As an example we show map of adaptive scales for the image "Lena" in Figure 2a. Note that the values of h^+_{ref} are maximal for large smooth regions, where distortions would be most visible. On contrary, the values of h^+_{ref} are approaching minimum on close-to-edge regions.

The structural similarity of two images can be evaluated by comparing the structural information contained in their adaptive scales. We introduce a *Window Term* (WT), which is a difference map of the adaptive scales for two images. The WT is defined in the following way:

$$\mathrm{WT}(x) = 1 - \frac{\left|h^+(x) - h^+_{ref}(x)\right|}{\max\limits_{x \in X}\left(\left|h^+ - h^+_{ref}(x)\right|\right)}, \qquad \forall x \in X \qquad (2)$$

where h^+_{ref} and h^+ are the maps of the adaptive scales for the reference image, and the test image, respectively. The values of WT vary from 0 (maximum difference) to 1 (no difference in the adaptive scales). Figure 2b shows the *Window Term* between the original "Lena" image and the blurred "Lena" on Figure 1a.

3. ICI-BASED QUALITY ASSESSMENT

The WT (2) gives a pointwise estimate of the local structural similarity between two images. In order to obtain a quality index valid for the whole image domain, we propose to take its average, which we name *mean Window Term* (mWT).

As shown in the next section, the mWT alone is a sufficiently accurate indicator of the image quality. However, in some situations the adaptive scales are the same for visually different images – e.g. negative or intensity shift, as shown in Figure 3. In such cases, WT alone would not be sufficient for image quality evaluation.

For handling such structure-preserving impairments, we introduce an *Intensity Term* (IT),

$$\mathrm{IT}(x) = 1 - \left(\frac{I(x) - I_{ref}(x)}{\mathrm{D}}\right)^2, \qquad \forall x \in X \qquad (3)$$

where I_{ref} is the reference image, I is the test image, and D is the dynamic range of the images. The values of the *Intensity Term* are in the range of 0 (maximum difference) to 1 (no difference). Figure 2c shows the *Intensity Term*, between the original "Lena" image and the blurred "Lena" on Figure 1a.

We combine the *Window Term* and the *Intensity Term* in the compound term ICIQ (for "ICI Quality"):

$$\mathrm{ICIQ}(\mathrm{x}) = \mathrm{IT}(\mathrm{x}) \cdot \mathrm{WT}(\mathrm{x}), \qquad \forall x \in X \qquad (4)$$

The ICIQ between original and blurred "Lena" is shown in Figure 2d.

Finally, we define the quality index mICIQ as the mean value of ICIQ over the image domain. It is in the range from 0, meaning "lowest quality" to 1, meaning "no visual difference".

a) Negative, mWT=1, mIT=0.64 b) Mean shift, mWT=1, mIT=0.88

Fig. 3. Impairments where mWT is the same as in the original image, while the mIT is different

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

We have compared three quality indices – mICIQ, mWT and MSSIM against two databases with subjective opinion scores – LIVE R1 and LIVE R2. For calculation of the

TABLE I

COMPARISON OF IMAGE QUALITY ASSESSMENT MODELS OVER LIVE DATABASE RELEASES 1. MAE: MEAN ABSOLUTE ERROR, RMS: ROOT MEAN SQUARE ERROR

Model	MAE	RMS
mICIQ	4.67	6.23
mWT	4.84	6.45
UQI	7.76	9.90
MSSIM	3.95	5.62
Sarnoff	4.66	5.81
PSNR	6.53	8.45

All models are done using logistic regression.

TABLE II

COMPARISON OF IMAGE QUALITY ASSESSMENT MODELS OVER LIVE DATABASE RELEASE 2. MAE: MEAN ABSOLUTE ERROR, RMS: ROOT MEAN SQUARE ERROR

Model	MAE	RMS
mICIQ	5.52	7.27
mWT	5.58	7.36
MSSIM	5.32	6.92

All models are using logistic regression.

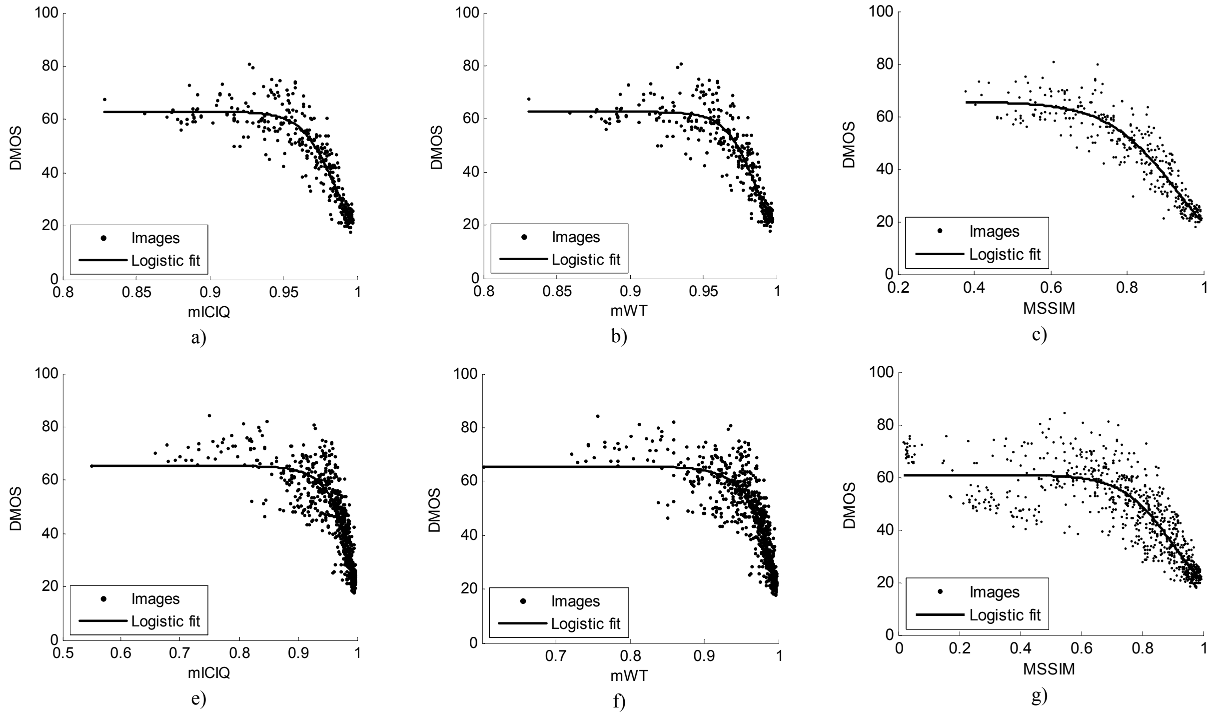


Fig. 3. Scatter plots of subjective distortion mean opinion score (DMOS) versus model prediction. Lower DMOS score means better quality.
a) mICIQ for LIVE database R1, b) mWT for LIVE database R1, c) MSSIM for LIVE database R1
d) mICIQ for LIVE database R2, e) mWT for LIVE database R2, f) MSSIM for LIVE database R2

adaptive scales we used $\Gamma = 30$ and $H = \{3, \cdots, 99\}$.

LIVE R1 and LIVE R2 are databases containing distorted images and their subjective evaluations [7]. LIVE database R1 consists of 460 images with JPEG and JPEG2000 artifacts. LIVE database R2 consists of 982 images – the images from R1, and additionally ones with Gaussian blur, white noise and artifacts caused by errors in JPEG2000 bit stream. In Figure 4 are shown the scatter plots and the results of non-linear regression fitting of logistic curves over the distortion mean opinion scores (DMOS) versus prediction models for LIVE R1 and R2 databases.

The numerical results for LIVE database R1 and R2 are shown in Table I and Table II, respectively. From the tables is seen that while our quality index gives more appropriate scores than MSSIM for some image artifacts (see Figure 1), it also performs comparably over a broader set of image impairments.

However, the most significant result is that a quality indicator solely based on the adaptive scales (mWT) is producing satisfactory results, demonstrating that similarity of adaptive scales is essential for satisfactory modeling of human quality perception.

The simple *Intensity Term*, introduced in order to handle the cases of structure-preserving distortions, improves the quality index, but not significantly. Future work should improve the Intensity Term to better model the intensity perception of the human vision.

The use of color information is also being considered.

REFERENCES

[1] Wang, Z., 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.

[2] Wang, Z., E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment”, Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, Pacific Grove, California, USA, Nov. 2003.

[3] Wandell, B.A., Foundations of vision, Sinauer Associates, Inc, Sunderland, Massachusetts, USA, 1995.

[4] Mannos, J.L., and D.J. Sakrison, “The Effects of a Visual Fidelity Criterion on the Encoding of Images”', IEEE Transactions on Information Theory, pp. 525-535, Vol. 20, No. 4, 1974.

[5] Ahumada, A.J., “Computational Image Quality Metrics: A Review”, Digest of SID International Symposium 1993, vol. 24, Playa del Rey, California, USA, pp. 305-308, 1993.

[6] Katkovnik, V., K. Egiazarian, and J. Astola, “Adaptive window size image de-noising based on intersection of confidence intervals (ICI) rule”, J. of Math. Imaging and Vision, vol. 16, no. 3, pp. 223-235, 2002.

[7] Sheikh, H.R., Z. Wang, L. Cormack, and A. C. Bovik, “LIVE Image Quality Assessment Database Release 2”, http://live.ece.utexas.edu/research/quality.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, “The SSIM Index for Image Quality Assessment, Matlab demo code” http://www.cns.nyu.edu/~zwang/files/research/ssim/index.html

[9] Goldenshluger, A., and A. Nemirovski, “On spatial adaptive estimation of nonparametric regression”, Math. Meth. Statistics, vol. 6, pp. 135-170, 1997.