CALIBRATING MS-SSIM FOR COMPRESSION DISTORTIONS USING MLDS

C. Charrier[1]*, K. Knoblauch*[2]*, L. T. Maloney*[3] *and A. C. Bovik*[4]

[1] University of Caen-Basse Normandie , GREYC, UMR CNRS 6072, Caen, France
[2] INSERM, U846, Stem Cell and Brain Research Institute, Bron, France
[3] New-York University, Department of Psychology, Center for Neural Science, NY, USA
[4] University of Texas at Austin, LIVE lab, Austin, TX, USA.

ABSTRACT

In this paper, we describe a recently developed method for assessing perceived image quality, Maximum Likelihood Difference Scaling (MLDS), and use it to assess the performance of MS-SSIM on compression distored images. MLDS allows us to quantify supra-threshold perceptual differences between pairs of images and to examine how perceived image quality, estimated through MLDS, changes as the compression rate is increased. We show how the data collected by MLDS allows us to recalibrate MS-SSIM to improve its performance.

Index Terms— Difference scaling, Genetic algorithm, MS-SSIM.

1. INTRODUCTION

Lossy image compression techniques such as JPEG2000 allow high compression rates, but only at the cost of perceived degradation in image quality. There is a considerable literature concerning how human observers perceive compression-induced degradation in images and how well Image Quality Assessment (IQA) algorithms predict human judgments of reduction in image quality as a function of compression.

The most commonly employed means to assess human judgment of image quality is to ask human observers to rate image quality directly on a numerical scale. Human judgments are ordinarily expressed as the Mean Opinion Score (MOS) obtained from a sufficiently large set of human observer ratings relative to a normalized scale defined by the International Telecommunications Union (ITU) [1].

The typical summary of the agreement between rated subjective image quality and the output of an IQA algorithm is some measure of the correlation between the subjective ratings and the measured degree of distortion. Typical measures of correlation include 1) Pearson's linear correlation coefficient (CC) between MOS and algorithm score after nonlinear regression, 2) the root-mean-squared error (RMSE) between MOS and the algorithm score after nonlinear regression and 3) the Spearman rank order correlation coefficient (SROCC).

This research is supported by the ANR project #ANR-08-SECU-007-04, and by the Intel and Cisco Inc under the VAWN program.

Among well-known IQA algorithms, Multi-Scale Structural SIMilarity (MS-SSIM) [2] computes relative quality scores between a reference image and a distorted version, achieving excellent correlations with MOS values.

Despite its success, MS-SSIM contains a number of parameter values that have not been optimized and remain somewhat ad hoc. Towards improving MS-SSIM, we used Maximum Likelihood Difference Scaling to investigate the manner in which algorithm scores vary from human scores, to guide the selection of parameter for better predicting compression distortions. The use of MLDS is justify by its ability to quantify supra-threshold perceptual differences between pairs of images and to examine how perceived image quality changes as the compression rate is increased to optimize the construction of psychovisual scale. Such a scale will serve as groundtruth to apply the parameter selection process.

2. MAXIMUM LIKELIHOOD DIFFERENCE SCALING

MLDS can be used to estimate the effect of compression on perceived image quality for any choice of image compression algorithm. In this section we explain the model of the observer's judgments in the psychophysical task on which MLDS is based.

An *image series* consist of a *base image* ϕ_1 and compressed versions of the base image denoted $\phi_2, \ldots, \phi_N$ numbered in increasing order of compression. If image ϕ_i is compressed to a greater degree than image ϕ_j we write $\phi_i < \phi_j$. For brevity we denote image in the series by their subscripts. The pair (i, j) will serve as shorthand for (ϕ_i, ϕ_j).

On each trial, the observer views two pairs of stimuli (i, j) and (k, l)representing four different levels of compression of the initial image (including possibly no compression). We refer to these two pairs as a *quadruple* denoted $\{i, j; k, l\}$. The observer judges whether the perceptual difference between the first pair (a, b) is greater than that between the second pair (c, d). Over the course of the experiment, he judges the differences of a subset of all possible quadruples (pairs of pairs) for the N stimuli in the series $\phi_1, ..., \phi_N$.

The goal of MLDS is to assign numerical scale values $(\psi_1, \psi_2, \ldots, \psi_N)$ that can be used to predict how the observer orders the pairs in each quadruple. We refer to these values as a *difference scale*. This difference scale can be effectively created if the observer satisfies three conditions [3] that are:

an ordering task that is a transitivity criterium :

$$(a_i \succ_1 a_j) \,\&\, (a_j \succ_1 a_k) \Rightarrow (a_i \succ_1 a_k).$$

where $\succ_1$ represents "is judged more distorted than", in terms of the observer perception. This first task yields a ranking along an axis based on a common property of stimuli.

an interval task that is also referred to the six points condition:

$$\left.\begin{array}{l}(a_i : a_j) \succ_2 (a_l : a_m) \\ \text{and} \\ (a_j : a_k) \succ_2 (a_m : a_n)\end{array}\right\} \Rightarrow (a_i : a_k) \succ_2 (a_l : a_n)$$

where $\succ_2$ represents the "perceived greater difference than" judgment. This criteria can be considered as transitivity of interval judgments.

a technical axioms task. these axioms yield the definition of a relationship between the set of stimuli (a_i) and the numerical values (n_i). Thus, there are numbers (n_i) such that:

$$\begin{array}{l}a_i \succ_1 a_j \Leftrightarrow n_i > n_j, \\ \text{and} \\ (a_i : a_j) \succ_2 (a_k : a_l) \Leftrightarrow \|n_i - n_j\| > \|n_k - n_l\|.\end{array}$$

Maloney and Yang [4] proposed a method to estimate the scale values by direct maximization of the likelihood. However, because the decision rule involves a simple linear combination of the internal responses, the scale values may also be estimated using a Generalized Linear Model (GLM) [5].

3. THE MS-SSIM ALGORITHM

As mentioned previously, the MS-SSIM index [6] is based on three multiscale factors: 1) the luminance distortion (ld) 2) the contrast distortion (cd) and 3) the structure distortion (sd) between an image f and a degraded version of it g.

From its basic formulation at any scale i, the luminance distortion is defined as

$$l(f,g) = \frac{2\mu_f\mu_g + C_1}{\mu_f^2 + \mu_g^2 + C_1} \tag{1}$$

where μ_f and μ_g represent the mean intensity of f and g at scale i, and C_1 is a constant to avoid instability when $\mu_f^2 + \mu_g^2 \approx 0$.

Contrast distortion at scale i is defined in a similar way *i.e.*:

$$c_i(f,g) = \frac{2\sigma_f\sigma_g + C_2}{\sigma_f^2 + \sigma_g^2 + C_2} \tag{2}$$

where C_2 is a non negative constant and σ_f (resp. σ_g) represents the standard deviation of f (and g) at scale i.

The structure comparison is performed after luminance subtraction and contrast normalization. The structure comparison function is defined as:

$$s_i(f,g) = \frac{2\sigma_{f,g} + C_3}{\sigma_f^2\sigma_g^2 + C_3} \tag{3}$$

where $\sigma_{f,g} = \frac{1}{N-1}\sum_{i=1}^{N}(f_i - \mu_f)(g_i - \mu_g)$, and C_3 is a small constant. Note that $sd(f,g)$ takes negative values whenever the local image structure is inverted.

Finally, The MS-SSIM value is computed by combining the luminance comparison (1), the contrast distortion measure (2) and the structure comparison (3) at different scales by

$$\text{MS-SSIM}(I,J) = [l_M(I,J)]^{\alpha_M} \prod_{i=1}^{M} [c_i(I,J))]^{\beta_i} [s_i(I,J))]^{\gamma_i} \tag{4}$$

where, the luminance comparison $ld_M(f,g)$ is computed only at scale M. The three exponents α_M, β_i and γ_i are used to adjust the relative importance of different components. In this paper, $M = 5$ corresponds to the maximum scale, while $i = 1$ corresponds to the original resolution of the image. In [2], the authors have defined $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$, and $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$.

4. CORRELATION STORY

We applied MLDS to evaluate the image quality of the 15 trial original images, each compressed with JPEG2000 to nine different levels: 0.1000, 0.3057, 0.5627, 0.7684, 0.9741, 1.1798, 1.3854, 1.5912 bpp, plus the original image. We used the JPEG2000 implementation provided by The JasPer Project. We obtained difference scales for each subject and image. In order to compare MLDS values with scores obtained from the MS-SSIM IQA algorithm, we computed the score provided by the IQA algorithm between consecutive pairs of compressed images, then cumulated these paired scores across the series.

In [7], the authors have found that even if MS-SSIM globally yields high correlations with the judgment of human observers, sometimes it fails to accurately predict perceptual changes as the compression rate is increased. More precisely, the third factor was less well correlated with MLDS than the two other factors, especially at the beginning of each scale. In order to counterbalance this lack of fit, a basic weighting rule that consists of modifying the weight value on the third factor has been investigated [7]. It has been found that refining the exponents values for the third MS-SSIM factor $s(.,.)$, the individual failure observed at the beginning of the scale tends to disappear, while the rest of the curve is unaffected, yielding a higher correlation value with human ratings. From this, it can

be presumed that to improve the correlation of the MS-SSIM IQA algorithm scores and MLDS, the coefficients (β_i, γ_i) do not necessarily have to be identical. Thus, next we investigate the impact of letting all of the parameters, α_i,β_i and γ_i, vary.

5. GENETIC OPTIMIZATION

5.1. The associated error function

The main objective is to find new exponent values for each decomposition scale of MS-SSIM. The associated formula can be expressed as a 15-parameter function :

$$\text{MS-SSIM}(I, J, \alpha_i, \beta_i, \gamma_i; i = 1, \cdots, M) = \prod_{i=1}^{M} \left[l_i(I,J)^{\alpha_i} c_i(I,J)^{\beta_i} s_i(I,J)^{\gamma_i} \right] \quad (5)$$

where $\sum_{i=1}^{M} \alpha_i + \beta_i + \gamma_i = 1$ and $\forall i \in [1, \cdots, M], 0 \leq \alpha_i \leq 1, 0 \leq \beta_i \leq 1, 0 \leq \gamma_i \leq 1$.

From (5), the search for the new exponent values seeks minimization of the error function

$$E(\alpha_i, \beta_i, \gamma_i; i = 1, \cdots, M) = \min \left(\sum_{j=1}^{K} (\text{MLDS}_j(I,J) - \text{fSSIM}_j(I, J, \alpha_i, \beta_i, \gamma_i))^2 \right) \quad (6)$$

where K is the number of tested images for which the MLDS values are provided, and $\text{fSSIM}_j(.)$ are the MS-SSIM computed rates obtained following a logistic regression as depicted in [9].

In other words, the goal is to estimate the 15 exponent values that minimize the error function $E(.)$. Since the error function is non-convex and may contain numerous local optima, the choice of search strategy to optimize it is important.

5.2. Search strategy

The Genetic Algorithm (GA) is a population-based stochastic search procedure that finds exact or approximate solutions to optimization and search problems. Modeled on the mechanisms of evolution and natural genetics, genetic algorithms uses directed random searches to locate optimal solutions in multimodal landscapes. Their basic principles were first introduced by Holland in 1975 [8].

Usually, a simple GA is composed of three operations: selection, genetic operation, and replacement. GAs use a population, which is composed of a group of chromosomes, to represent the solutions of the system. Defining the solution representation of the system is the first task when applying GAs. The solution in the problem domain can then be encoded into the chromosome in the GA domain, and *vice versa*. Initially, a population is randomly generated. The fitting function then uses values from objective functions to evaluate the quality of fit of each chromosome. Next, a particular group of chromosomes is chosen from the population to be parents. The offspring are then generated from these parents using genetic operations (crossover and mutation). The fitness of the offspring are then evaluated and used in replacement processes in order to replace the chromosomes in the current population by the selected off-spring. The GA cycle is then repeated until a desired termination criterion is satisfied, or the objective value is below the threshold.

In this paper, $M = 5$ is the number of levels used to compute the MS-SSIM value. In that case, the GA domain represents a 15-dimensional space in which one point is expressed as $(\alpha_1, \cdots, \alpha_M, \beta_1, \cdots, \beta_M, \gamma_1, \cdots, \gamma_M)$, and the fitness function is defined by (6).

5.3. Optimization results

Table 1 shows the estimated values for each exponent after minimizing (6). In addition, confidence intervals with a 95% confidence level are provided for each exponent. They are computed using a bootstrap process with 999 replicates. If we consider the associated coefficients for the structure attribute (γ values), we observe that the third decomposition level seems to be of greater importance since its exponent value γ_3 is higher whereas the four others are quite similar.

6. EVALUATION OF THE PERFORMANCE OF THE REFINED MS-SSIM INDEX

In order to judge the relevance of the 15 new exponents estimated in the previous section, we tested the refined MS-SSIM index on the LIVE Image Quality database.

To provide quantitative performance evaluation, three measures of correlation have been used: 1) Pearson, 2) Kendall and 3) Spearman measures. To perform the Pearson correlation measures, a logistic function (as adopted in the video quality experts group (VQEG) Phase I FR-TV test [9]) was used to provide a non-linear mapping between the refined MS-SSIM values and subjective scores. We then separately used the subjective scores provided with the overall LIVE database. Kendall and Spearman correlation measures were computed between the DMOS values and the MS-SSIM indices obtained using both the original exponent values and the new ones (Table 1).

Considering the LIVE database, the results are presented in Table 2 where bold face values represent statistical significant difference corresponding to p-value inferior to 0.5. From correlation evaluation results, we see that the performance of the MS-SSIM index computed with the new exponent values yields improved performance relative to the MS-SSIM values obtained with the original exponent values. This is not true for noisy or blurred images, since a decrease of the correlation coefficients is observed. Nevertheless, when all degradations are included, one observes that the SROCC is significantly higher when new exponent values are used. Naturally, this is

Exponent	α_1	α_2	α_3	α_4	α_5
Value	0.1920	0.2169	0.2026	0.2136	0.1749
CI	[0.0989,0.2415]	[0.1877,0.2791]	[0.1692,0.2384]	[0.1765,0.2868]	[0.0814,0.2304]
Exponent	β_1	β_2	β_3	β_4	β_5
Value	0.9612	0.0097	0.0097	0.0097	0.0097
CI	[0.8288,0.9681]	[-0.0145,0.0933]	[0.0084,0.0112]	[0.0084,0.0112]	[-0.0133,0.1012]
Exponent	γ_1	γ_2	γ_3	γ_4	γ_5
Value	0.0082	0.1586	0.8167	0.0083	0.0082
CI	[0.0073,0.0086]	[0.1241,0.2530]	[0.7250,0.8501]	[0.0073,0.0086]	[0.0073,0.0086]

Table 1. The 15 computed exponents and associated confidence intervals (CI) with a 95% confidence level using a GA approach.

	JP2K		JPEG	
	Original	New	Original	New
CC	0.783	0.810	0.730	0.742
KROCC	0.884	0.884	0.849	0.852
SROCC	0.980	0.991	0.962	0.981
	Gaussian blur		FastFading	
	Original	New	Original	New
CC	0.8864	0.8623	0.725	0.788
KROCC	0.8591	0.8413	0.859	0.876
SROCC	0.9725	0.9627	0.965	0.974
	White Noise		All	
	Original	New	Original	New
CC	0.9153	0.9142	0.7980	0.8142
KROCC	0.8887	0.8878	0.8021	0.8543
SROCC	0.9825	0.9813	0.9464	**0.9762**

Table 2. Computed correlation coefficients between original MS-SSIM indices and MS-SSIM indices using the new exponents based on the analysis described here. Statistically significant differences (with p-value inferior to 0.5) are bold face.

driven in part by optimization of QA with respect to JP2K and also FastFading (which uses JP2K), but also JPEG distortion.

7. CONCLUSION

We have used a novel psychophysical method, Maximum Likelihood Difference Scaling (MLDS), to address the limitations inherent in the MS-SSIM IQA method. We applied it to a large collection of images to assess the consequences of JP2K compression and compared observers' judgments image quality to the MS-SSIM predictions. We found that MS-SSIM suffers from local failures when assessing JP2K compression, especially due to its structure factor that greatly influences the predicted values. It was found that these local failures can be reduced using different values for the three $(\alpha_i, \beta_i, \gamma_i)$ exponents which we estimate from data. The refined MS-SSIM index yielded significantly improved performance relative to the original algorithm.

8. REFERENCES

[1] ITU-R Recommendation BT.500-11, "Méthodologie d'évaluation subjective de la qualité des images de télévision," Tech. Rep., ITU, Geneva, Switzerland, 2002.

[2] 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*, 2003, pp. 1398–1402.

[3] J. N. Yang and L. T. Maloney, "Difference scaling in color space near the neutral point," in *Investigative Ophtalmology and Visual Science*, Fort Lauderdale, Florida, May 1998, vol. 39 of *ARVO annual meeting*, p. 160, abstracts.

[4] L. T. Maloney and J. N. Yang, "Maximum likelihood difference scaling," *Journal of Vision*, , no. 3, pp. 573–585, 2003.

[5] C. Charrier, Laurence T. Maloney, H. Cherifi, and K. Knoblauch, "Maximum likelihood difference scaling of image quality in compression-degraded images," *Journal of the Optical Society of America*, vol. 24, no. 11, pp. 3418–3426, 2007.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, 2004.

[7] C. Charrier, K. Knoblauch, A. K. Moorthy, A. C. Bovik, and L. T. Maloney, "Comparison of image quality assessment algorithms on compressed images," in *SPIE,Image Quality and System Performance VII*, San-Jose, California, Jan. 2010.

[8] John H. Holland, *Adaptation in natural and artificial systems*, MIT Press, Cambridge, MA, USA, 1992.

[9] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Tech. Rep., 2000.