Comparison of Image Quality Assessment Algorithms on Compressed Images

Christophe Charrier[1], Kenneth Knoblauch[2], Anush K. Moorthy[3]
Alan C. Bovik[3], Laurence T. Maloney[4]

[1] Université de Caen-Basse Normandie, GREYC UMR CNRS 6072, Equipe Image
[2] INSERM, U846, Stem Cell and Brain Research Institute, Département Neurosciences Intégratives, Bron, France
[3] University of Texas at Austin, LIVE lab
[4] Department of Psychology, Center for Neural Science, New York University

ABSTRACT

A crucial step in image compression is the evaluation of its performance, and more precisely the available way to measure the final quality of the compressed image. Usually, to measure performance, some measure of the covariation between the subjective ratings and the degree of compression is performed between rated image quality and algorithm. Nevertheless, local variations are not well taken into account.

We use the recently introduced Maximum Likelihood Difference Scaling (MLDS) method to quantify supra-threshold perceptual differences between pairs of images and examine how perceived image quality estimated through MLDS changes the compression rate is increased. This approach circumvents the limitations inherent to subjective rating methods.

Keywords: Maximum-Likelihood difference scaling, quality assessment

1. INTRODUCTION

Lossy image compression techniques such as JPEG2000 allow high compression rates, but only at the cost of some perceived degradation in image quality. Image quality is ultimately measured by how human observers react to compressed/degraded images and many evaluation methods are based on eliciting ratings of perceived image quality. One of the challenges induced by the use of such a compression method is the measure of the resulting quality. The steps in an image quality assessment (IQA) algorithms typically consist of 1) performing a color space transformation to obtain decorrelated color coordinates and 2) decomposing these new coordinates into perceptual channels. An error is then estimated for each of these channels. The final quality score is obtained by pooling these errors in the spatial and/or frequenty domain.

Formally, the main goal of IQA algorithms is to predict ratings as done by human beings. To quantify human judgments, the Mean Opinion Score (MOS) is obtained from a set of human observer ratings with respect to a normalized scale defined by the International Telecommunications Union (ITU).[1]

The typical summary of the agreement between rated image quality and algorithm is some measure of the covariation between the subjective ratings and the degree of compression. Typical measures of covariation 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). Ultimately, however, the interpretation of human ratings is difficult. Suppose, for example, that the human observer rates two compressed images as 3 and 4 in image quality and also rates two other images as 7 and 8, respectively. Although the difference in rating is the same

Further author information: (Send correspondence to C.C.)
C.C.: E-mail: christophe.charrier@unicaen.fr
K.K.: E-mail: ken.knoblauch@inserm.fr
A.K.M.: E-mail: anushmoorthy@gmail.com
A.C.B.: E-mail: bovik@ece.utexas.edu
L.T.M.: E-mail: ltm1@nyu.edu

Image Quality and System Performance VII, edited by Susan P. Farnand, Frans Gaykema, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7529, 75290B · © 2010 SPIE-IS&T · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.840221

for both pairs, we have no way to conclude whether the perceived increase in quality between the first pair of images is equal to, greater than, or less than, the perceived increase in quality between the second pair. The subjective ratings only allow us to order the images by quality.

In this paper, we evaluate the efficacy of recently-developed general-purpose image quality assessment (IQA) algorithms for assessing the compression-quality tradeoff.

CHARRIER *et al.*[2] recently applied a novel psychophysical method, Maximum Likelihood Difference Scaling (MLDS)[3,4] that circumvents the limitations inherent to subjective rating methods. MLDS allows us to quantify supra-threshold perceptual differences between pairs of images in order to evaluate the perceptual changes in images as the compression rate is increased. The MLDS method is based on simple, forced-choice judgments and requires remarkably few trials to obtain quantitative estimates of the effects of any degree of image compression.[2]

This paper is structured as follows. In Section 2, we present the MLDS method. Section 3 details the used experimental setup. In section 4, obtained results are presented and discussed. Last Section concludes.

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 $\{i, j; k, l\}$. He 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, the observer judges the differences of a subset of all possible quadruples (pairs of pairs) for the N stimuli in the series $\phi_1, ..., \phi_N$. (i.e., N compression levels).

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*. In principle, we wish to assign these scale values so that the perceived difference between the images of the pair (ϕ_i, ϕ_j) is judged greater than perceived difference between the images of the pair (ϕ_k, ϕ_l)if and only if,

$$\|\psi_i - \psi_j\| > \|\psi_k - \psi_l\|. \tag{1}$$

However, if the differences $\|\psi_i - \psi_j\|$ and $\|\psi_k - \psi_l\|$ are close it is unlikely that human observers would be so reliable in judgment as to satisfy the criterion just given in Eq. 1. To take into account this judgment variation, MALONEY and YANG[3] proposed a model of difference judgment that allows the observer to exhibit some stochastic variation in judgment. We next describe their model. Let $L_{ij} = \|\psi_i - \psi_j\|$ be the *length* of the interval (a_i, a, j). The proposed decision model is an equal-variance Gaussian signal detection model,[5] where the signal is the difference in the lengths of the intervals,

$$\delta(i, j; k, l) = L_{ij} - L_{kl} = \|\psi_i - \psi_j\| - \|\psi_k - \psi_l\| \tag{2}$$

The signal δ is contaminated by Gaussian error ϵ with mean 0 and standard deviation σ to form the judgment variable

$$\Delta(i, j; k, l) = \delta(i, j; k, l) + \epsilon. \tag{3}$$

MALONEY and YANG assumed that the observer, given the quadruple $(i, j; k, l)$, selects the pair (i, j) precisely when $\Delta(i, j; k, l) > 0$. The resulting model of the observer allows for stochastic variation in judgment. When the magnitude of $\delta(i, j; k, l)$ is small is small relative to the Gaussian standard deviation, σ, the observer, presented with the same stimuli, can give different responses. The degree of inconsistency predicted depends of the magnitude of $\delta(i, j; k, l)$ relative to σ and this dependence can be used to test the model itself.[3,4]

(a) imga (b) imgb (c) imgc (d) imgd (e) imge

(f) imgf (g) imgg (h) imgh (i) imgi (j) imgj

(k) imgk (l) imgl (m) imgm (n) imgn (o) imgo

Figure 1. The 15 images used in the experiments are shown, with mnemonic labels. For each image, we estimated a difference scale based on each observer's judgments, a total of 450 difference scales.

Let R_t be the observers response to the tth quadruple $(i_t, j_t; k_t, l_t)$ in the experiment, $t = 1, \ldots, n$. R_t is coded as follows: $R_t = 0$ if the difference of the first pair is judged to be larger, and $R_t = 1$ otherwise. We select the parameters $\Psi = (\psi_1, \psi_2, \ldots, \psi_N)$ and σ to maximize the likelihood:

$$L(\Psi, \sigma; X) = \prod_{t=1}^{n} \left[\Phi\left(\frac{\delta_t}{\sigma}\right)^{1-R_t} \left(1 - \Phi\left(\frac{\delta_t}{\sigma}\right)\right)^{R_t} \right] \tag{4}$$

where δ_t is given by Eq. 2 applied to the tth quadruple $(i_t, j_t; k_t, l_t)$, Φ represents the cumulative Gaussian distribution function, and R_t is the $t^{th}; t = 1, n$.

We estimated difference scales for each observer's data for each image, using MLDS as described above. We employed multiple starting points to minimize the possibility of encountering local minima. All computations were carried out in the statistical language R using the `optim` function. We have integrated the functions necessary to perform these fits using either approach in an R package (MLDS) available from the Comprehensive R Archive Network (CRAN, accessible from `http://www.rproject.org/`).

If we add a constant c to all the values on the difference scale $(\psi_1, \psi_2, \ldots, \psi_N)$ that maximizes likelihood, the resulting difference scale also maximizes likelihood. If we multiply all the values on the maximum likelihood difference scale $(\psi_1, \psi_2, \ldots, \psi_N)$ by a positive constant $a > 0$ the resulting difference scale also maximizes likelihood once we scale σ by a. Therefore, without loss of generality, we can fix the end points of the maximum likelihood difference scale to be $\psi_1 = 0$ and $\psi_N = 1$. We report all our results in this normalized format.

3. EXPERIMENTAL SETUP

3.1 Apparatus

Thirty observers participated to the psychophysical tests. All observers had normal color vision (Ishihara test) and normal or corrected-to-normal acuity (Snellen test).

We computed 15 image series using the base images shown in Fig. 1. These images portray a variety of scenes and differ in distribution of spatial and chromatic detail.

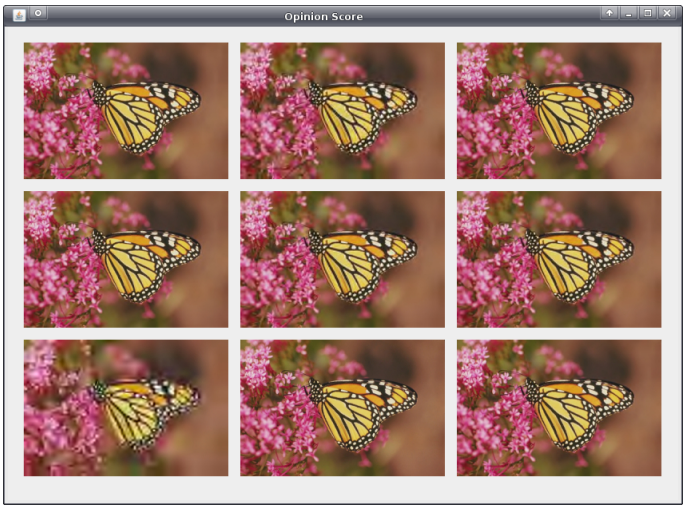


Figure 2. Example of a single trial during the ordering test. The subject sees an image at the nine trial compression rates. The stimuli are ramdomly arranged on 3 lines. The subject was asked to order the quality of all images from the best to the worst quality.

We first tested whether observers could correctly order images in descending order of quality. If they cannot do so, allowing for possible difficulty in discriminating adjacent images in the scale, there is no difference scale that can account for their performance.

During this initial test, observers had to first select the highest quality image, then the second highest, etc. A sample trial is shown in Fig. 2. During the test,each time an image is selected by clicking on it, the selected image disappears and the number of the rank order is shown. If the observer decided to cancel his choice, he just has to click on the rank order number. The corresponding image is shown and the rank order number disappears. In addition, he can deselect more than one image, depending on the selected number. for example, if the observer has already classified six images, the observer can deselect any image numbered from 1 to 6. If he deselects image numbered 3, all images from 3 to 6 will be automatically deselected.

During the second psychophysical task, the observer saw a quadruple of images drawn from a single image series. These four images were arranged as two pairs (i, j) and (k, l) on a computer display. On half the trials, the first pair was displayed on the upper half of the display screen, the second on the lower, and on the remaining trials the first pair was displayed on the lower, the second on the upper. For the convenience of the observer, the less compressed of the two images in each pair was always on the left. The observer then must judge which pair (upper or lower))exhibited the larger change or difference in quality. A sample trial is shown in Figure 3. Over the course of the experiment, the observer judged several hundred quadruples. These judgments can be used to construct a numerical difference scale that captures the effect of additional compression on image quality.

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 obtained difference scales for each subject and image.

In order to compare MLDS values with scores obtained from an IQA algorithm, we computed the score provided by the IQA algorithm between consecutive pairs of compressed images and cumulated these paired scores across the series. The MLDS values were obtained by first recording the judgments of human observers in a psychophysical experiment in which they classified between which pair of 2 pairs of images was the perceived difference greatest. The scale values that best predicted their judgments were estimated by a maximum likelihood procedure[3, 4]

To perform the comparison, a prior-normalization process of the obtained objective quality scores is needed to be able to fit those values the MLDS ones.

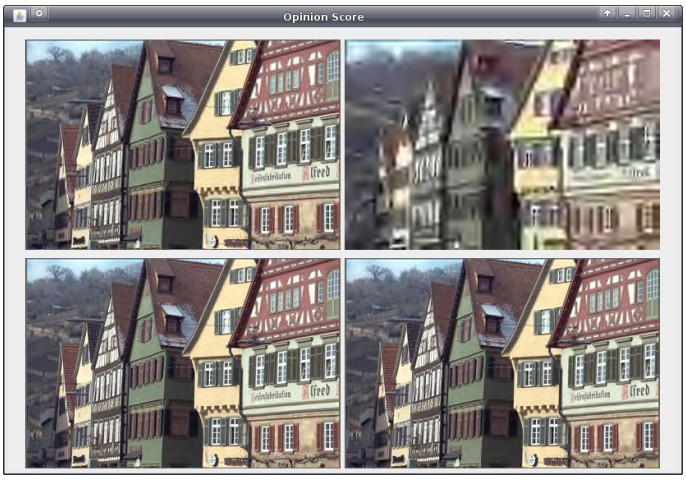


Figure 3. Example of a single trial in MLDS. the subject sees an image at four different compression rates. The stimuli are arranged as two pairs (i, j) and (k, l). In each pair, the second stimulus is more compressed. The subject was asked to judge whether the decrease in quality in going from i to j is greater than the decrease in going from k to l. In this example, most observers would judge that the upper pair exhibits the larger change.

3.2 the IQA algorithm

The MS-SSIM index, is a multiscale extension of the SSIM IQA algorithm. The MS-SSIM index proposed by Wang,Simoncelli and Bovik[6] contains three factors: 1) a luminance distortion, 2) a contrast distortion and 3) a structure comparison. The philosophy of this measure lies in the representation of an image by a vector within an image space. Thus, any image distortion can be interpreted as adding a distortion vector to the reference image vector. In this space, the two vectors that represent luminance and contrast changes span a plane that is adapted to the reference image vector. The image distortion corresponding to a rotation of such a plane by an angle can be interpreted as the structural change.

The luminance comparison is defined as

$$l(I, J) = \frac{2\mu_I\mu_J + C_1}{\mu_I^2 + \mu_J^2 + C_1} \tag{5}$$

where μ_I and μ_J respectively represent the mean intensity of the image I and J, and C_1 is a constant avoiding instability when $\mu_I^2 + \mu_J^2 \approx 0$. According to the Weber's law, the magnitude of a just-noticeable luminance change δL is proportional to the background luminance L. In that case, $\mu_I = \alpha\mu_J$, where α represents the ratio of the luminance of the distorted signal relative to the reference one. The luminance comparison can be now defined as

$$l(I, J) = \frac{2\alpha\mu_I^2 + C_1}{(1 + \alpha^2)\mu_I^2 + C_1} \tag{6}$$

The contrast distortion measure is defined in a similar form:

$$cd(I, J) = \frac{2\sigma_I\sigma_J + C_2}{\sigma_I^2 + \sigma_J^2 + C_2} \tag{7}$$

where C_2 is a non negative constant, and σ_I (resp. σ_J) represents the standard deviation .

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

$$s(I, J) = \frac{2\sigma_{I,J} + C_3}{\sigma_I^2\sigma_J^2 + C_3} \tag{8}$$

where $\sigma_{IJ} = \frac{1}{N-1}\sum_{i=1}^{N}(I_i - \mu_i)(J_i - \mu_J)$, and C_3 is a small constant. $s(I, J)$ can take negative values which is interpreted as local image structure inversion.

Finally the MS-SSIM measure is given by:

$$\text{MS-SSIM}(I,J) = [l_M(I,J)]^{\alpha_M} \prod_{i=1}^{M} [cd_i(I,J))]^{\beta_i} [s_i(I,J))]^{\gamma_i} \tag{9}$$

where, the contrast comparison and the structure comparison are computed at the i^{th} scale, and denoted as $cd_i(I,J)$ and $s_i(I,J)$, respectively; the luminance comparison $l_M(I,J)$ 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. $M = 5$ corresponds to the maximum scale, while $i = 1$ corresponds to the original resolution of the image. In,[6] the authors have defined $\alpha_M = 1$ and $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$, and $\beta_5 = \gamma_5 = 0.1333$.

4. RESULTS AND DISCUSSION

The obtained results (Fig. 4) show that MS-SSIM captures perceptual changes in images with increasing compression rates very well. Yet, even if MS-SSIM globally yields high correlations with the judgment of human observers, sometimes this IQA algorithm fails to accurately predict perceptual changes in images as the compression rate is increased for particular images.

In order to investigate these local failures, the same procedure that was used to compare the scores obtained from the IQA algorithm and MLDS values is used for each one of the three factors embedded within the MS-SSIM index. The results are shown in Fig. 5, where the first row of each subimage corresponds to the luminance comparison values $l_M(I,J)$, the second row correponds to the contrast comparison values $\prod_{i=1}^{M} cd_i(I,J)$, and the last row represents the structure comparison values $\prod_{i=1}^{M} s_i(I,J)$. At first glance, one can remark that the third factor is not as well correlated with MLDS, especially at the beginning of the scale. In addition, the same kind of curve is noticeable with MS-SSIM. Thus, structure comparison $\prod_{i=1}^{M} s_i(I,J)$ is of great influence on the MS-SSIM values, as it has been suggested by.[7]

To reach the best fitting as possible, one has to reduce the influence of this third parameter. This is performed in two ways: 1) investigation of the influence of the decomposition level M and 2) the addition of a weight term to this third parameter.

As the third parameter is initially computed using $M = 5$ levels, we first investigate the influence of that decomposition level: how M influences the curve of this third factor. To measure this influence, we compute the structure comparison factor for levels from 1 to 5. The obtained results are shown in Fig. 6, where the black points and the blue curve respectively represent the MLDS and the third factor values.

Generally, the blue curve best fits MLDS when the decomposition level increases. In addition, one can observe that either $M = 4$ or $M = 5$ gives the best global results.

Yet, for each decomposition level, one can observe a severe lack of good fit at the beginning of each scale, for each trial image. This lack of fitting is obviously observed for low decomposition level values ($M = 2$, $M = 3$). The higher the decomposition level, the less the lack of fitting is.

In order to counterbalance this lack of fitting, one investigate a weighting rule that could be applied to the third parameter. As this weight is mainly dedicated to better fit the beginning of the associated curve to the MLDS values, this weight can be interpreted as a parameter slope κ. In that case, Eq. 9 would be reformulated as follows

$$\text{MS-SSIM}(I,J) = [l_M(I,J)]^{\alpha_M} \prod_{i=1}^{M} [cd_i(I,J))]^{\beta_i} \left[\prod_{i=1}^{M} [s_i(I,J))]^{\gamma_i} \right]^{\kappa}. \tag{10}$$

with $0 \leq \kappa \leq 1$.

To find the best slope parameter value κ, three measure of the covariation between MLDS and the weighted MS-SSIM values are performed: Pearson's linear correlation coefficient (CC) between MLDS and algorithm score after nonlinear regression, 2) the Kendall rank order correlation coefficient (KROCC) between MLDS and the algorithm score and 3) the Spearman rank order correlation coefficient (SROCC). Fig. 7 presents the obtained

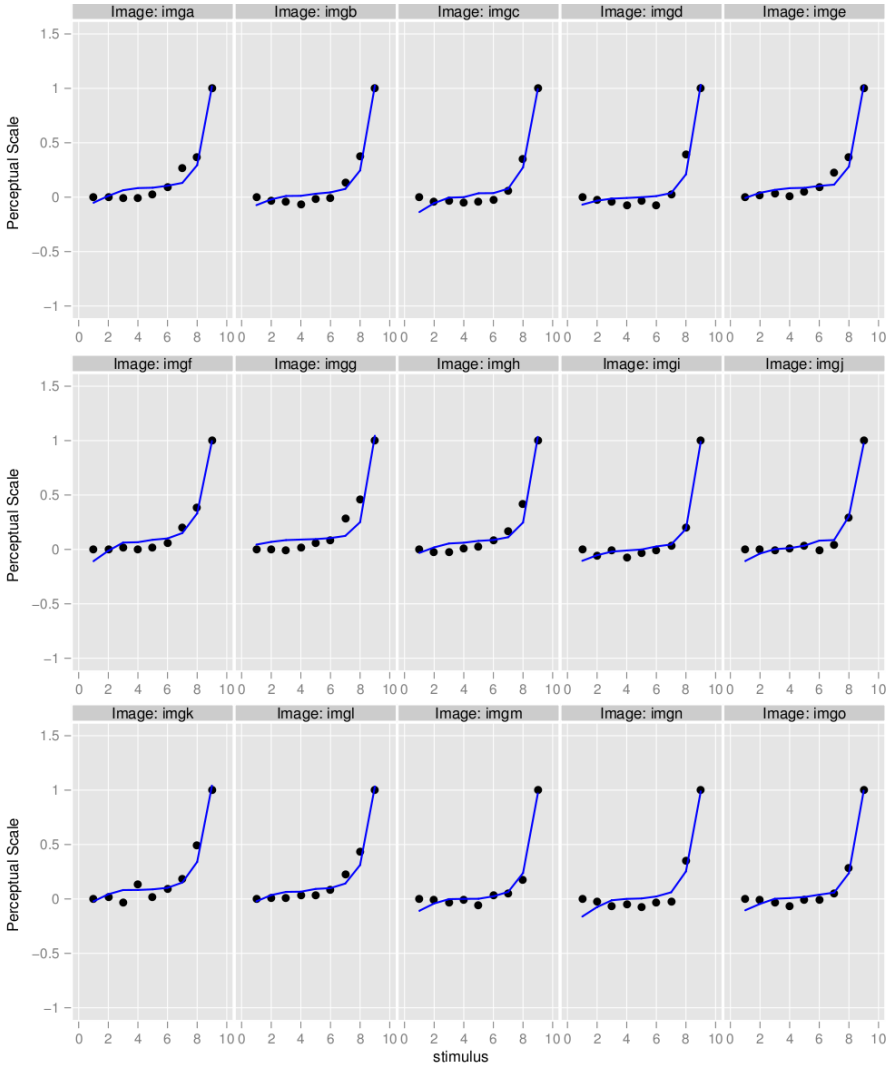


Figure 4. Obtained results for all trial images. The black points and the blue curve respectively represent the MLDS and the MS-SSIM values.

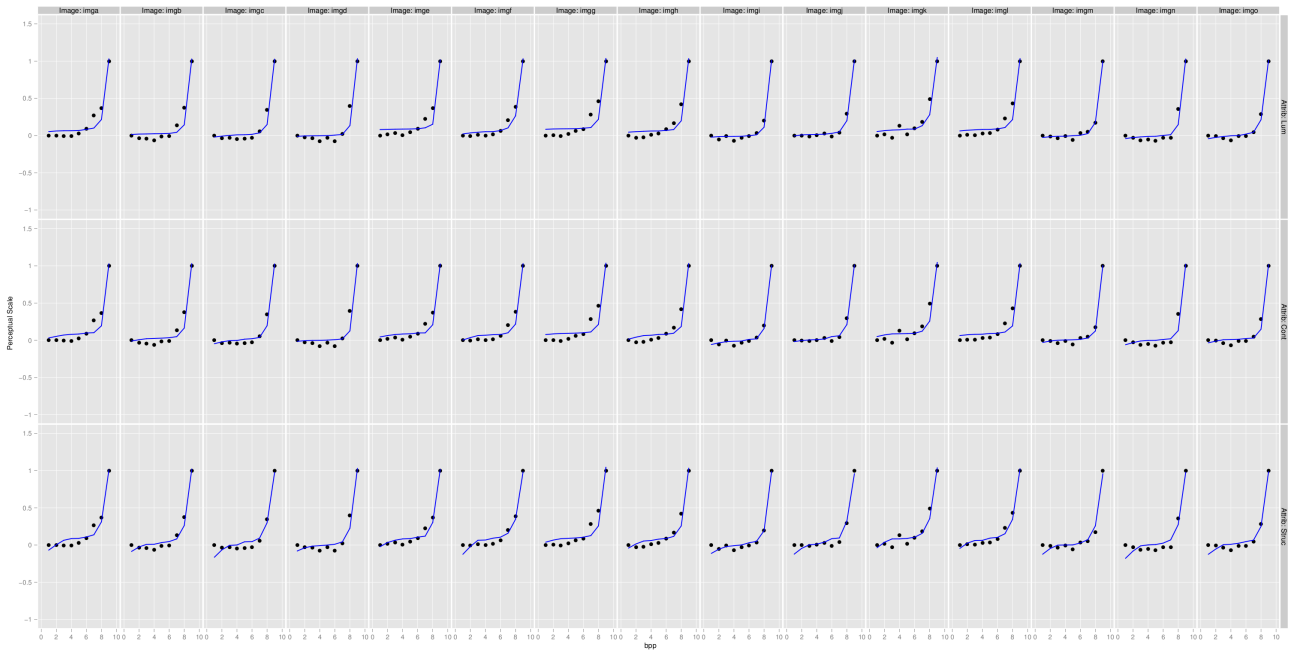


Figure 5. Obtained results for all trial images. The black points and the blue curve respectively represent the MLDS and each of the three MS-SSIM factors values. The first row of each subimage corresponds to the luminance comparison values $l_M(I, J)$, the second row corresponds to the contrast comparison values $\prod_{i=1}^{M} cd_i(I, J)$, and the last row represents the structure comparison values $\prod_{i=1}^{M} s_i(I, J)$

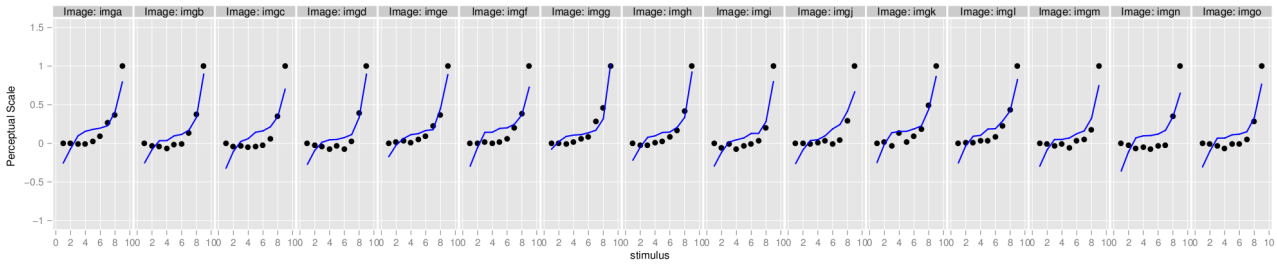
results, where the blue line represents the correlation coefficients computed between MLDS the original MS-SSIM algorithm score, and the red line corresponds to the correlation coefficients computed between MLDS the weighted MS-SSIM algorithm score. From Fig. 7(b) and Fig. 7(c), the maximum correlation values correspond to a weight $\kappa = 0.14$.

From initial MS-SSIM IQA algorithm (Eq. 9), we saw that the three initial exponents are equal $\alpha_i = \beta_i = \gamma_i$ and $\sum_{i=1}^{M} = 1$. Using such coefficients allows us to reach a high degree of correlation between MOS and the IQA metric.[6] In that case, the MS-SSIM is considered as globally excellent (but not perfect since correlation coefficients values differ from 1) to predict quality scores. In fact, that could be interpreted in two ways. The first one consists of considering an exact prediction global failure; in that case, all predicted scores are very close to the actual scores but a small residual error exists for all scores. The second way consists of a quasi exact prediction of actual values (residual errors are very small) except for local failures. Considering Fig. 4, the situation seems to be the second one, where a local failure is highlighted at the beginning of the scale.

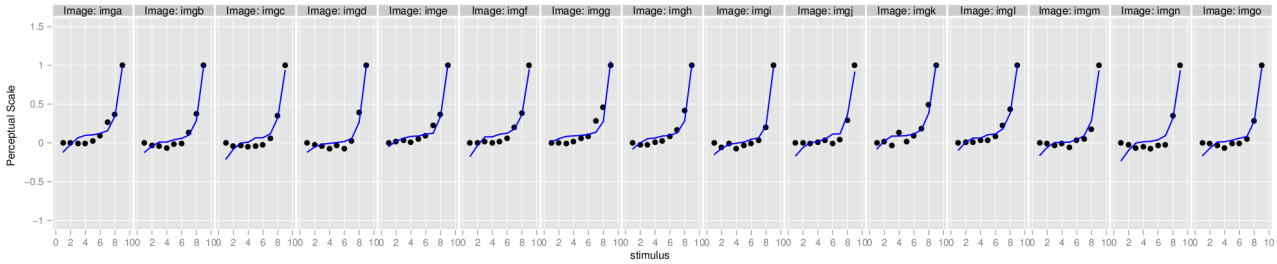
When using a value of κ that differs from 1, one observes that, for the structure comparison factor the local failure at the beginning of the scale tends to disappear whereas the global curve aspect remains. Fig. 8) illustrates that point of view, where the blue line corresponds to the initial structure comparison attribute, and the green line corresponds to the structure comparison attribute with a weight $\kappa = 0.14$.

This shows that to improve the correlation of the MS-SSIM IQA algorithm scores and MLDS, all implied coefficients $(\alpha_i, \beta_i, \gamma_i)$ do not have to be identical (as initially suggested).

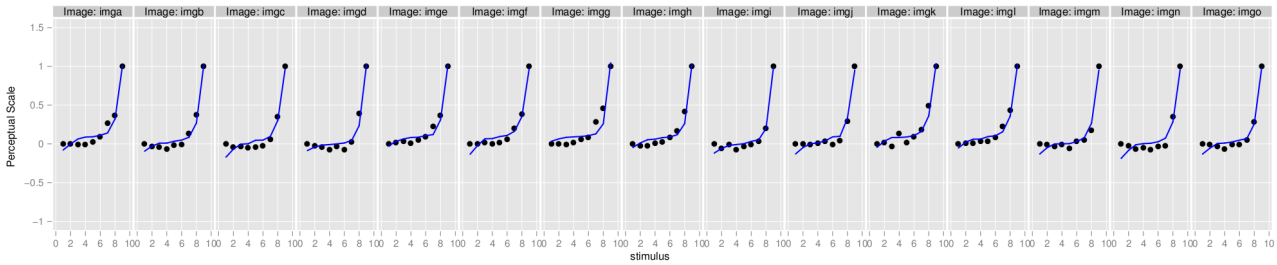
The MLDS allows us to find that, even if MS-SSIM the IQA algorithm shows a high correlation degree with MLDS, it suffers from local failures. These local failures are mainly due to the structure comparison factor that is of great influence on the predicted scores. These local failures could be corrected by changing the coefficients γ_i.



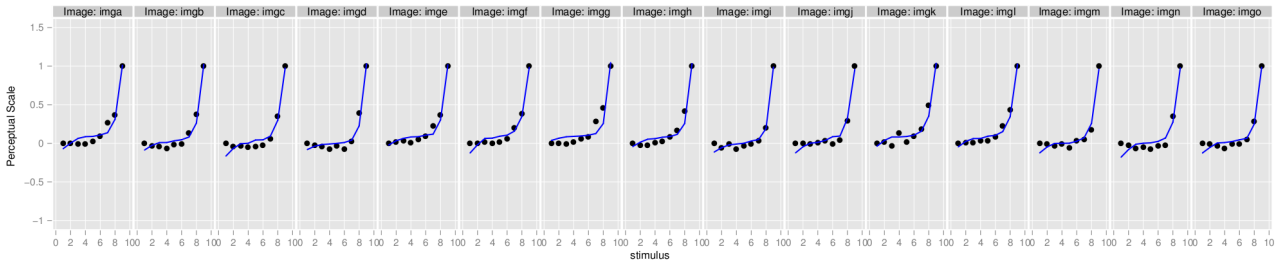
(a) Structure comparison (Eq. 8) computed with a decomposition level M=2



(b) Structure comparison (Eq. 8) computed with a decomposition level M=3

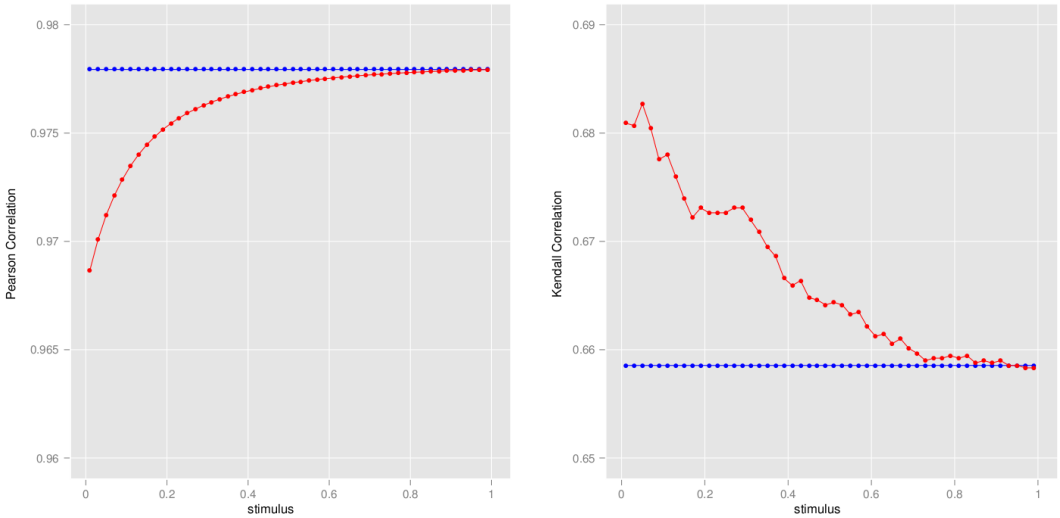


(c) Structure comparison (Eq. 8) computed with a decomposition level M=4



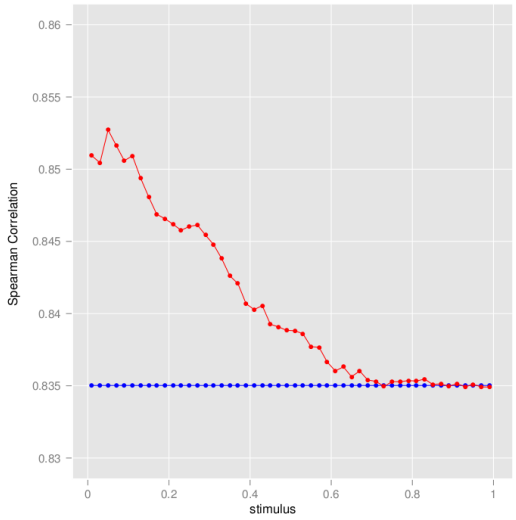
(d) Structure comparison (Eq. 8) computed with a decomposition level M=5

Figure 6. The structure comparison feature values (Eq. 8) as used to compute the MS-SSIM values, for different decomposition levels. The black points and the blue curve respectively represent the MLDS and the third factor values used to compute MS-SSIM.



(a) Pearson correlation coefficient

(b) Kendall correlation coefficient



(c) Spearman correlation coefficient

Figure 7. Obtained results for the three commonly used correlation coefficient: Pearson, Kendall and Spearman. The red line and the blue line respectively represent the computed correlation coefficient between MLDS and the original MS-SSIM values and between MLDS and weighted MS-SSIM.

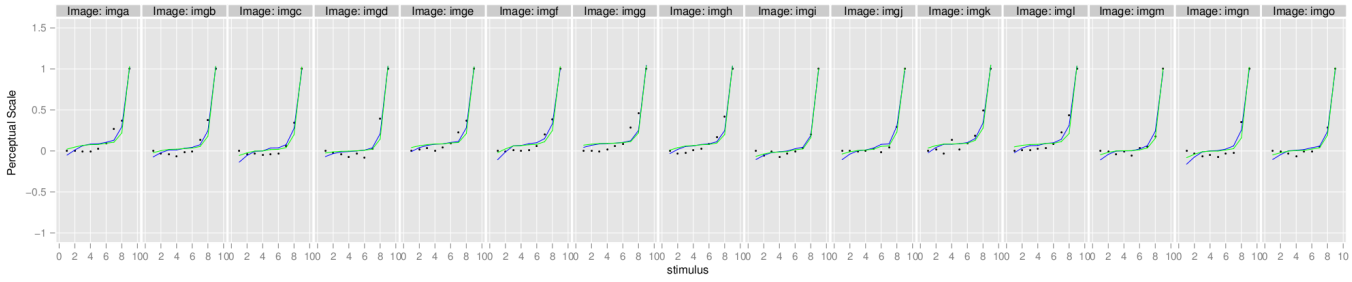


Figure 8. The structure comparison factor for all images. The blue line corresponds to the initial structure comparison attribute, and the green line corresponds to the structure comparison attribute with a weight $\kappa = 0.14$

5. CONCLUSION

In this paper, we have used a novel psychophysical method, Maximum Likelihood Difference Scaling (MLDS) to circumvent the limitations inherent to subjective rating methods. Indeed, to measure the performance of any IQA algorithm, one usually uses some correlation coefficients. Yet, in such case, it is impossible to understand to what a failure is due (a global or a local problem). Using MLDS, one are able to know the type of failure : global or local.

Applying this method to MS-SSIM IQA technique, one shows that it suffers from local failures especially due to its third factor that greatly influences the predicted values. These local failures can be reduced using different values for the three $(\alpha_i, \beta_i, \gamma_i)$ exponents.

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. C. Charrier, L. 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* **24**(11), pp. 3418–3426, 2007.
3. L. T. Maloney and J. N. Yang, "Maximum likelihood difference scaling," *Journal of Vision* (3), pp. 573–585, 2003.
4. K. Knoblauch and L. T. Maloney, "MLDS: Maximum likelihood difference scaling in R," *Journal of Statistical Software* **25**, pp. 1–26, 1 2008.
5. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*, Robert E. Krieger Publishing Company, 1974.
6. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, pp. 1398–1402, 2003.
7. D. M. Rouse and S. S. Hemami, "Understanding and simplifying the structural similarity metric," in *International Conference on Image Processing (ICIP'08)*, pp. 1188–1191, 2008.