

# Analysis of the Difference of Gaussians Model in Image Difference Metrics

Sebastien A. Ajagamelle<sup>1,2</sup>, Marius Pedersen<sup>2,3</sup>, Gabriele Simone<sup>2</sup>;

1: Grenoble Institute of Technology; Grenoble, France.

2: Gjøvik University College; Gjøvik, Norway.

3: Océ Print Logic Technologies S.A., France.

## Abstract

*The goal of this work is to present and review two new image difference metrics, named  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$ . These metrics are along the same lines as the standard S-CIELAB metric (Zhang and Wandell, 1997), modified to include a pyramidal subsampling, the Difference of Gaussians receptive-field model (DOG) (Tadmor and Tolhurst, 2000), and the  $\Delta E_E$  color-difference formula (Oleari et al., 2009). The DOG model and the  $\Delta E_E$  formula have been shown to improve respectively contrast measures and image quality metrics (Simone et al., 2009). Extensive testing using 29 state-of-the-art metrics and six image databases has been performed. Although this new approach is promising, we only find weak evidence of effectiveness. Analysis of the results indicates that the metrics show fairly good correlations over particular test images, yet they do not outperform the most common objective quality measures.*

## Introduction

In the field of digital imaging, technology advancements are rapid, and new methods are frequently proposed to deal with the limitations of various imaging systems. Images are subject to a wide variety of distortions, such as compression artifacts and noise. For this reason, reliable quality measurement is needed to evaluate the quality of digital images. Subjective rating by human observers has been the most precise way to measure Image Quality (IQ), but it is often time and resource demanding. Objective evaluation has de facto been proposed. In this paper, we focus on one objective method to measure IQ, IQ metrics. An IQ metric usually aims to mathematically predict human visual perception. Although some metrics closely simulate this perception, scientific research has not yet been able to provide a carbon copy of the Human Visual System (HVS). This accounts for the outstanding number of IQ metrics that have been developed so far [1], and not all of them show consistency with subjective evaluation. The yardstick to judge the performance of a metric is a combination of simplicity, modularity, prediction accuracy, and low computational cost. Because of the computational complexity and the multifaceted aspects of IQ, we focus on Image Difference (ID) which is the first step to be able to calculate IQ [2].

We will only consider full-reference IQ metrics, i.e., metrics assuming that an original image is available. These bivariate quantitative measures usually follow the same general framework. The original image and its reproduction are first transformed into a suitable color space, preferably a perceptually uniform one. Then a simulation of the HVS is carried out, from simplistic methods, as smoothing the image by a local neighborhood, to more intricate methods, as using Contrast Sensitivity Functions (CSFs). In most cases, these metrics

eventually perform a calculation of difference using a color-difference formula. We put forward two new ID metrics inspired from the S-CIELAB framework [3], though built upon a DOG filtering, which has been shown to improve contrast measures [4, 5].

First, an insight into the state of the art will be provided, followed by a description of the proposed metrics and the common IQ metrics selected to make a comprehensive comparison. We will assess the performance of the IQ metrics across several image databases, and subsequently analyze and discuss the results. Finally we conclude and propose future work.

## State of the Art

The CIELAB color space specification is a modified version of Adam's chromatic value diagram using a nonlinear transformation of the CIE XYZ tristimuli. It was initially developed for color patches to provide a perceptually uniform color space and it adopts the Euclidean distance, thus providing a computationally simple way to measure color differences. When dealing with digital images, the  $\Delta E_{ab}^*$  formula is usually utilized to calculate the color difference between a reference image and a sample image in each pixel separately.

In 1997, Zhang and Wandell heralded a new era in objective IQ assessment by introducing a spatial extension to the CIELAB color difference named S-CIELAB [3], that was laid out because  $\Delta E_{ab}^*$  does not correlate properly with perceived IQ. The main thrust of their research was to offer both a spatial filtering to simulate the blurring of the HVS and a consistency with  $\Delta E_{ab}^*$  for large uniform areas. In the S-CIELAB framework, the first step is a separation of the images into an opponent-color space (achromatic, red-green, and blue-yellow space). Each opponent color image is then convolved with a two-dimensional spatial kernel determined by the visual spatial sensitivity of that color dimension. The original S-CIELAB uses unit sum kernels to filter the independent channels. The filtered images are finally transformed back to CIE XYZ and then to CIELAB, and pixelwise difference are computed using the CIELAB formulae. S-CIELAB is often thought of as a benchmark metric because of its implementation simplicity and because it has inspired many other IQ metrics, such as SHAME [6],  $S - CIELAB_{JOHNSON}$  [7], and S-DEE [8].

Latterly, the  $\Delta E_E$  color-difference formula has been published [9]. It is a Euclidean formula that was initially devised for small-medium color differences in the log-compressed OSA-UCS space. The OSA-UCS color-appearance system was developed by the Committee on Uniform Color Scales of the Optical Society of America whose purpose was to realize a space of ceramic tile colors with perceived uniform color scales. The investigation on the color opponencies showed that the OSA-UCS color space has a particular structure that can be closely

related to a possible model of the human visual processing [10, 11].

Contrast is an important image attribute that falls under the umbrella of IQ [12]. It is linked to saturation and spatial resolution, since higher saturation will usually increase the variance of the image, and higher spatial resolution will enable an observer to see more details. Measuring contrast is a challenging assignment. Many parameters, such as viewing distance, light conditions, image content, memory color, experience, or the contextual dependence on the observer task, affect how observers perceive the scene. The historical milestones in the search for a characterization of this attribute consist of global measures, which assume that the response of the HVS depends much less on the absolute luminance than on the relation of its local variations. These global measures are usually not suitable for measuring perceived contrast of real visual configurations since they exhibit many shortcomings, one of them being that a few points of extreme brightness or darkness can determine the contrast of the whole image [13].

Tadmor and Tolhurst [4] developed a local contrast measure based on the DOG receptive-field model, modified and adapted to natural images. The conventional model describes the spatial sensitivity in the center of receptive fields (central component) by a bi-dimensional Gaussian with a peak amplitude at 1.0:

$$Center(x,y) = \exp \left[ - \left( \frac{x}{r_c} \right)^2 - \left( \frac{y}{r_c} \right)^2 \right],$$

where  $x$  and  $y$  indicate the row and the column of the pixel  $(x,y)$ , and the radius  $r_c$  represents the distance beyond which the sensitivity declines following  $1/e$  regarding to the peak level.

The surround component is represented by a Gaussian curve as well, with a larger radius  $r_s$ :

$$Surround(x,y) = 0.85 \left( \frac{r_c}{r_s} \right)^2 \exp \left[ - \left( \frac{x}{r_s} \right)^2 - \left( \frac{y}{r_s} \right)^2 \right],$$

the scaling factor  $0.85 \left( \frac{r_c}{r_s} \right)^2$  fixes the integrated sensitivity of the surround component so that it is equal to 85% of that of the center. For a central point of the receptive-field positioned at  $(x,y)$ , the output of the center component for an image pixel at position  $(i,j)$  is given by:

$$R_c = \sum_i \sum_j Center(i-x, j-y) Picture(i, j),$$

while the output of the surround component is:

$$R_s = \sum_i \sum_j Surround(i-x, j-y) Picture(i, j).$$

The conventional model assumes that the response of a neurone is only determined by local luminance differences between the receptive-field center and surround:

$$DOG(x,y) = R_c(x,y) - R_s(x,y).$$

Three criteria for the measure of contrast were proposed, where the response gain is set by the local mean luminance:

$$C(x,y) = \frac{R_c(x,y) - R_s(x,y)}{R_c(x,y)},$$

$$C(x,y) = \frac{R_c(x,y) - R_s(x,y)}{R_s(x,y)},$$

$$C(x,y) = \frac{R_c(x,y) - R_s(x,y)}{R_c(x,y) + R_s(x,y)}.$$

In 2004 Rizzi et al. proposed the RAMMG contrast measure [14]. In this measure, a transformation from  $RGB$  to the CIELAB color space is first applied, followed by a pyramidal subsampling of the image to various levels. A calculation of the local contrast in each pixel is carried out by taking the average of absolute value difference between the lightness channel value of the pixel and the surrounding eight pixels, resulting in a contrast map for each level separately. The final overall measure is a recombination of the average contrast for each level.

Following the similar approach introduced in RAMMG, Rizzi et al. proposed in 2008 the Retinal-like Subsampling Contrast algorithm (RSC) [5]. The RSC works with the same pyramid subsampling as RAMMG, but it computes for each pixel of each level the DOG contrast calculation, and this computation is performed separately for the lightness and the chromatic channels. The three measures are then combined with different weights.

Simone et al. [15] developed shortly afterwards a weighted level framework (WLF) as an evolution of the previous contrast measures. The main improvements are the use of antialiasing filtering in the pyramid construction combined with a weighted recombination of the local contrast maps. The measure can be extended to different color spaces and is not limited to CIELAB, as RSC and RAMMG.

## From Contrast to Image Difference : Two New Metrics

Two new metrics are proposed referred as  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$ . They are in line with the S-CIELAB approach, but the spatial extension is based on the work initiated by Rizzi et al. [14] and refined by Simone et al. [8]. The improvements encompass a multi-level approach, the substitution of the original S-CIELAB spatial filtering with a DOG calculation, and the use of the  $\Delta E_E$  color-difference formula. The general workflow of the metrics is as follows (Figure 1):

- The original image and its reproduction are first converted into the CIELAB color space for  $S_{DOG} - CIELAB$  and into CIE XYZ for  $S_{DOG} - DEE$ .
- The images are subsampled to various levels afterwards. The undersampling is simple since the images are halved, and the antialiasing filtering avoids artifacts at low resolutions.
- A pixelwise neighborhood contrast calculation is executed in each level using the DOG on the lightness and on the chromatic channels separately, thus providing local contrast maps for each level and each channel.
- Local contrast errors are computed using  $\Delta E_{ab}^*$  for  $S_{DOG} - CIELAB$  or  $\Delta E_E$  for  $S_{DOG} - DEE$ .
- A weighted recombination of the local contrast maps is finally computed, resulting in global ID metrics.

These metrics grew out of research on contrast measures, therefore one could expect the highest correlations with images containing significant variations in contrast. The DOG model was chosen as a surrogate to the CSF filtering because it has revealed to be beneficial in the identification of edges, while the CSFs are mainly used to modulate the less perceptible frequencies. The  $\Delta E_E$  formula was selected because it is statistically equivalent to CIEDE2000 in the prediction of many available empirical datasets, but with greater simplicity and clear relationships with visual processing.

Once that local contrast maps are generated for each level, how to reduce the concept of contrast from local values at each pixel location to a single number representing the global ID is an ongoing debate. The simplest strategy is taking the mean of each level and averaging all together.

These two new metrics perform a weighted recombination of the levels, given by the following equation:

$$GlobalID = \frac{1}{N_l} \sum_{l=1}^{N_l} \lambda_l \cdot \bar{c}_l,$$

where  $N_l$  is the number of levels,  $\bar{c}_l$  is the mean contrast in the level  $l$ , and  $\lambda_l$  is the weight assigned to each level  $l$ .

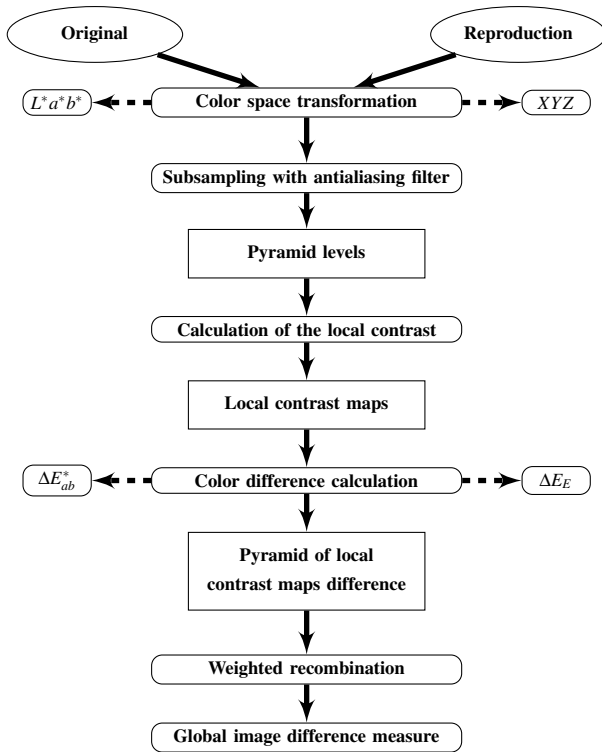


Figure 1: Workflow of the proposed metrics. The metrics are similar to the S-CIELAB described by Zhang and Wandell. The improvements encompass a multi-level approach, the substitution of the S-CIELAB spatial filtering with a DOG calculation, and the use of the  $\Delta E_E$  color-difference formula.

We selected several parameters in order to see whether a particular configuration would result in better adequacy with subjective evaluation. Table 1 describes the different values of the central component and surround component of the DOG filter, the nature of the subsampling, and the type of weighting levels.

Parameter	Set 1	Set 2	Set 3	Set 4	Set 5
$r_c$ (pixel)	1	1	2	3	2
$r_s$ (pixel)	2	2	3	4	4
Type of pyramid	P1	P1	P1	P1	P1
Type of weighting levels	1:1	Var.	Var.	Var.	Var.

Table 1: Set of parameters for  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$ . The way of building the pyramid is expressed by the series  $P_1 = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \dots$ . The variance (Var.) was used to give importance to high-resolution levels

The variance was used for the weighting level parameters so as to give importance to high-resolution levels, and because it could be a possible key to improve perceptual ID metrics, as it has been proven for measuring contrast [15].

For a detailed overview of the parameters, we refer the reader to Simone et al. [13].

## Evaluation of the Proposed Metrics

We carried out an extensive evaluation of the new metrics by selecting a multitude of test images. Among the few public databases providing images for evaluation of IQ metrics, we used the Tampere Image Database 2008 (TID2008) [16] and the IVC image database [17], together with four datasets containing respectively luminance changed images [18, 19], JPEG and JPEG2000 compressed images [20], images with global variations of contrast, lightness, and saturation [21], and gamut mapped images [22, 23].

The performance in correlation for each metric is calculated by a comparison between the perceptual difference issued from psychophysical experiments and the difference calculated by the metric. We opted for three standard types of correlation:

- The product-moment correlation coefficient or Pearson's correlation coefficient, which assumes a normal distribution in the uncertainty of the data values and that the variables are ordinal.
- The Spearman's rank-correlation coefficient, which is a non-parametric measure of association based on the ranks of the data values, that describes the relationship between the variables without making any assumptions about the frequency distribution.
- The Kendall's tau-rank correlation coefficient, which is a non-parametric test used to measure the strength of the dependence between the variables being compared when the data are in ordinal form.

In addition to the metrics previously introduced, we also compare with:

**S-CIELAB-JOHNSON** [7]: The  $S - CIELAB_{JOHNSON}$  metric works as the traditional S-CIELAB, but the spatial filters have been improved with different CSFs.

**S-DEE** [8]: This metric, proposed by Simone et al., is a modification of the S-CIELAB from Johnson, where the  $\Delta E_{ab}^*$  color-difference formula is replaced with the  $\Delta E_E$  formula.

**Hue angle algorithm** [24, 25]: This algorithm is based on the fact that systematic errors over the entire image are quite noticeable and unacceptable. The metric is based on the hue histogram and uses the  $\Delta E_{ab}^*$  as color-difference formula.

**SHAME and SHAME-II** [6]: The Spatial Hue Angle Metric (SHAME) proposed by Pedersen et al. can be considered as the combination of the original S-CIELAB and the original hue angle algorithm. SHAME-II is a variation of SHAME that applies the filtering used by Johnson and Fairchild [7].

**Universal Image Quality** [26]: The Universal Image Quality (UIQ) is a mathematically defined IQ metric for grayscale images. The index models any distortion as a combination of loss of correlation, luminance distortion, and contrast distortion.

**Structural similarity** [27]: The Structural Similarity (SSIM) index is based on the UIQ. It defines the structural information in an image as those attributes that represent the structure of the objects in the scene, independently of the average luminance and contrast. The comparisons are done for local windows in the image, and the overall IQ is the mean of all these windows.

**SSIM-IPT** [28]: This color extension of the previous metric calculates the SSIM for each channel of the IPT color space. Then it combines all the three channels using the geometrical mean.

**Qcolor** [29]: This color image fidelity metric is also derived from the UIQ. The UIQ is performed on each channel in the  $l$ ,  $\alpha$ , and  $\beta$  channels, that result from a transformation of the LMS space. The overall quality is a combination of the results for the three channels.

**Visual Signal-to-Noise Ratio** [30]: The Visual Signal-to-Noise Ratio (VSNR) quantifies the visual fidelity of natural images based on near-threshold and suprathreshold properties of human vision by using wavelet-based models.

**Visual Information Fidelity** [31]: This metric is based on the HVS and uses an additive white Gaussian noise model. It quantifies the information content of the reproduction relative to the information in the original image. The natural scene model used is a Gaussian scale model in the wavelet domain.

**HVS REAL** [32]: This image similarity metric is based on a multiscale model of the HVS that includes different channels accounting for perceptual phenomena such as color, contrast, color-contrast, and orientation selectivity. Features are extracted from these channels to create an aggregate measure of similarity using a weighted linear combination of the feature differences.

**PSNR-HVS** [33]: This algorithm computes a modified version of the Peak-Signal-To-Noise ratio (PSNR), where a scanning window is used to remove mean shifting and contrast stretching in order to take into account spatial sensitivity of the HVS.

**PSNR-HVS-M** [34]: PSNR-HVS-M is an enhancement of PSNR-HVS that includes notably a different reduction by values of contrast masking determined by means of CSF.

**Adaptive Bilateral Filter** [35]: This metric uses a bilateral filter, which avoids the loss of edge information when smoothing, that was optimized to be adaptive to corresponding viewing conditions and image entropy.

We also selected standard color-difference formulae ( $\Delta E_{94}^*$  and  $\Delta E_{00}^*$ ) and numerical objective quality measures (**MSE**, **RMS**, **PSNR**, **structural content**, **average difference**, **N-cross correlation**, **correlation quality**, **maximum difference**, and

**image fidelity**) [36, 37]. These latter are straightforward metrics designed to quantify the closeness of the altered and reference image fields and thus do not take into consideration the viewing conditions or any feature of the HVS.

For an overview of these IQ metrics and others, we refer the reader to Pedersen and Hardeberg [1].

### Evaluation Using the TID2008 Database

The TID2008 database contains a total of 1700 images, with 25 reference images and 17 types of distortions over 4 distortion levels (Figure 2 and Table 2). The mean opinion scores (MOS) are the results of 654 observers attending the experiments. No viewing distance is stated in the TID database, therefore we have used a standard viewing distance for the metrics requiring this setting.



Figure 2: The TID2008 database contains 25 reference images with 17 types of distortions over 4 levels.

	Type of distortion	Dataset							Full
		Noise	Noise2	Safe	Hard	Simple	Exotic	Exotic2	
1	Additive Gaussian noise	+	+	+	-	+	-	-	+
2	Noise in color components	-	+	-	-	-	-	-	+
3	Spatially correlated noise	+	+	+	+	-	-	-	+
4	Masked noise	-	+	-	+	-	-	-	+
5	High frequency noise	+	+	+	-	-	-	-	+
6	Impulse noise	+	+	+	-	-	-	-	+
7	Quantization noise	+	+	-	+	-	-	-	+
8	Gaussian blur	+	+	+	+	+	-	-	+
9	Image denoising	+	-	-	+	-	-	-	+
10	JPEG compression	-	-	+	-	+	-	-	+
11	JPEG2000 compression	-	-	+	-	+	-	-	+
12	JPEG transmission errors	-	-	-	+	-	-	+	+
13	JPEG2000 transmission errors	-	-	-	+	-	-	+	+
14	Non eccentricity pattern noise	-	-	-	+	-	+	+	+
15	Local block-wise distortion	-	-	-	-	-	+	+	+
16	Mean shift	-	-	-	-	-	+	+	+
17	Contrast change	-	-	-	-	-	+	+	+

Table 2: Overview of the distortions in the TID database and how they are related to the tested subsets. The database contains 17 types of distortions over 4 distortion levels. The sign "+" indicates that the distortion type was used to alter the images of the subset and the sign "-" that it was not considered for this subset.

TID2008 database

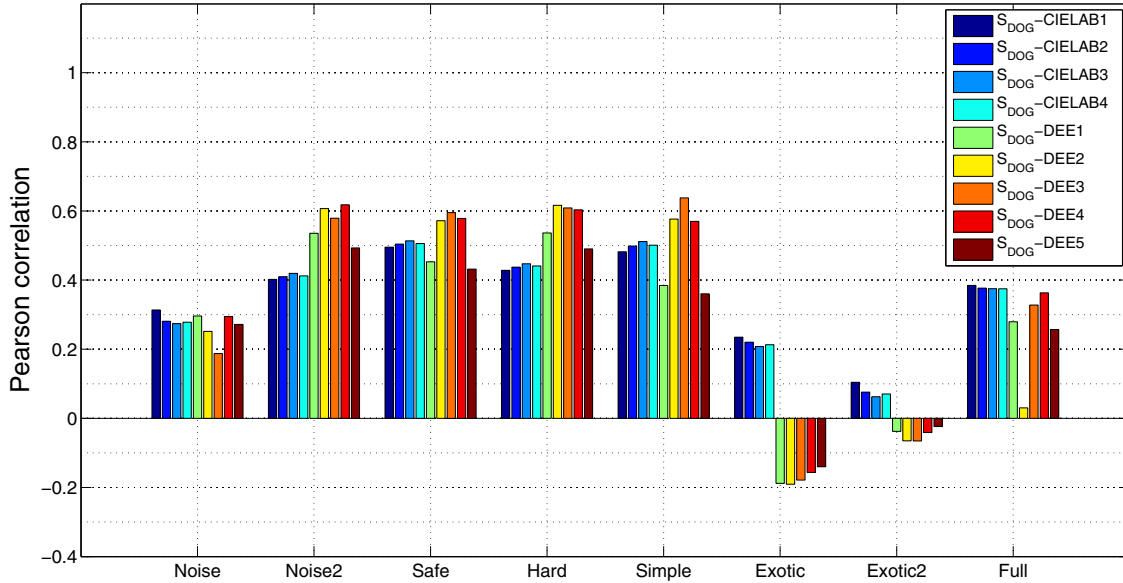


Figure 3: Pearson correlation coefficients for  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$  based on the TID2008 database. The correlations are higher for the datasets Noise2, Safe, Hard, and Simple. The Spearman and Kendall correlations follow the similar trend.

Figure 3 shows that the two proposed metrics have low performance over both the image datasets and the full database. We can see that there are differences in the correlations depending on the selected configuration. In particular,  $S_{DOG} - CIELAB$  shows the best correlations with the 1st set of parameters and  $S_{DOG} - DEE$  with the 4th set of parameters. We used only the first four sets of parameters to calculate  $S_{DOG} - CIELAB$  since the 5th configuration resulted in an unstable metric. We can see that the differences between the four correlations are small.

The correlations for all the IQ metrics are listed in Table 3. We found that  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$  are little better than simple color-difference formulae  $\Delta E_{ab}^*$  and  $\Delta E_E$ , as well as a few IQ metrics, such as S-DEE, SHAME,  $S - CIELAB_{JOHNSON}$ , and the hue angle algorithm. However, we see that all these correlations remain weak since they are under 0.5. Conversely, other metrics, such as UIQ, VIF, P-HVS and P-HVS-M, are more capable of predicting perceived image differences.

When looking at specific distortions (Table 2),  $S_{DOG} - CIELAB$  does not perform well. The highest Pearson correlation coefficients are 0.511 on the Simple dataset and 0.513 on the Safe dataset containing noise, blurring, and compression errors (Figure 3), with a best Pearson correlation for the full database of 0.385 (Table 3).  $S_{DOG} - DEE$  shows almost the same correlations as  $S_{DOG} - CIELAB$  for the full database, with a best Pearson correlation of 0.363. We found that  $S_{DOG} - DEE$  has higher correlation coefficients for the specific distortions Noise2, Safe, Hard, and Simple. This suggests that the metric is likely to be used to predict perceived image difference with images containing the distortions associated with these subsets. However,  $S_{DOG} - DEE$  cannot be reliably used as the sign of the correlation coefficient does not remain the same. For the other subsets Noise, Exotic, and Exotic2 containing denoising, pattern noise, local block-wise distortions, mean shift, and contrast change, we get low correlation coefficients indicating problems with the metric for these distortion types.

The poor results demonstrate that the proposed metrics are limited in their scope and performance, and therefore should be improved for the distortions found in the TID2008 database.

Metric	Pearson	Spearman	Kendall
$S_{DOG} - CIELAB1$	0.385	0.369	0.257
$S_{DOG} - CIELAB2$	0.377	0.356	0.246
$S_{DOG} - CIELAB3$	0.375	0.351	0.243
$S_{DOG} - CIELAB4$	0.374	0.353	0.244
$S_{DOG} - DEE1$	0.279	0.280	0.192
$S_{DOG} - DEE2$	0.342	0.350	0.244
$S_{DOG} - DEE3$	0.328	0.348	0.246
$S_{DOG} - DEE4$	0.363	0.364	0.253
$S_{DOG} - DEE5$	0.257	0.260	0.179
S-CIELAB	0.433	0.445	0.314
$\Delta E_{ab}^*$	0.232	0.276	0.186
$\Delta E_E$	0.273	0.381	0.266
$S - CIELAB_{JOHNSON}$	0.318	0.312	0.209
$S - DEE$	0.294	0.288	0.196
Hue angle	0.262	0.295	0.198
SHAME	0.300	0.351	0.246
SHAME-II	0.408	0.405	0.273
UIQ	0.622	0.596	0.433
SSIM	0.550	0.634	0.459
SSIM-IPT	0.484	0.570	0.406
$Q_{COLOR}$	0.522	0.482	0.338
VSNR	0.171	0.706	0.532
VIF	0.741	0.754	0.586
HVS REAL	0.199	0.175	0.120
PSNR-HVS	0.616	0.635	0.488
PSNR-HVS-M	0.589	0.606	0.466
ABF	0.275	0.259	0.175
$\Delta E_{94}^*$	0.086	0.242	0.165
$\Delta E_{00}^*$	0.086	0.243	0.165
MSE	0.535	0.561	0.406
RMS	0.536	0.561	0.406
PSNR	0.508	0.561	0.406
Structural Content	0.025	0.106	0.063
Average Difference	0.153	0.227	0.165
N-Cross-Correlation	0.087	0.347	0.238
Correlation Quality	0.026	0.045	0.031
Maximum Difference	0.189	0.452	0.315
Image Fidelity	0.253	0.532	0.378

Table 3: Comparison of the IQ metrics over all the images of the TID2008 database. The results show that UIQ, VIF, P-HVS, and P-HVS-M correlate rather well with subjective evaluation. The best correlation coefficients for  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$  demonstrate that the proposed metrics have fairly low performance for this set of test images.

## Evaluation Using Other Databases

We listed in Table 4 the Pearson correlation coefficients of the IQ metrics for these image databases. Figure 4 illustrates the performance of the metrics.

### Luminance changed images, Pedersen et al.

The database includes four original images reproduced with different changes in lightness. Each scene has been altered in four ways globally and four ways locally [18, 19].

For this database, only a handful of IQ metrics correlate reasonably well with subjective assessment. The original hue angle algorithm, SHAME-II, and ABF exhibit the best Pearson correlation coefficients. Additionally, we found that color-difference formulae ( $\Delta E_{ab}^*$ ,  $\Delta E_{94}^*$ , and  $\Delta E_{00}^*$ ) and trivial metrics (MSE and RMS) outperform  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$ . The conclusion is that the proposed metrics are not good estimates where the type of distortion is a local luminance shift. The results also indicate that, as for the images from the TID2008 database,  $S_{DOG} - CIELAB$  shows the best correlations with the 1st configuration and  $S_{DOG} - DEE$  with the 4th configuration. The differences between the configurations are minor, though.

### JPEG and JPEG2000 compressed images, Caracciolo et al.

The original images of this database were corrupted by JPEG and JPEG2000 distortions, generating a total of 80 degraded images [20]. The parameters for these distortions were randomly chosen with predefined ranges.

We found that all the metrics have low performance for the images from this database, spanning a range below 0.5. This is probably because these particular images were initially selected in order to determine the just noticeable distortion. Only small distortions were applied to the original images making it arduous for the observers to assess IQ, and therefore also very difficult for the IQ metrics. Regarding the performance of the proposed metrics, the best Spearman correlation coefficients are 0.232 for  $S_{DOG} - CIELAB1$  and 0.146 for  $S_{DOG} - DEE3$ . PSNR-HVS-M, PSNR, and maximum difference have slightly better performance than the other metrics.

### IVC database, Le Callet et al.

The IVC database contains blurred images and images distorted by three types of lossy compression techniques - JPEG, JPEG2000, and Locally Adaptive Resolution [17].

The Pearson correlations for this database are in general higher than those of the Caracciolo database, probably because the range of the distortions was selected differently. The most accurate IQ metrics are UIQ, VSNR, and VIF. When analyzing the results, we can see that  $S_{DOG} - DEE$  performs better than  $S_{DOG} - CIELAB$  and that  $S_{DOG} - DEE3$  provides the highest correlation coefficient.

### Images altered in contrast, lightness, and saturation, Ajagamelle et al.

This database contains a total of 10 original images covering a wide range of characteristics and scenes [21]. The images were modified on a global scale with separate and simultaneous variations of contrast, lightness, and saturation.

The results confirm that the proposed metrics work better with global variations of image attributes. Because of the DOG spatial filtering, they are also more efficient with contrast altered images.  $S_{DOG} - CIELAB$  has almost the same Pearson correlation coefficients as  $S_{DOG} - DEE$ , but both metrics are

slightly worse than most of the rest, indicating that the settings of the new metrics should be refined. We can notice the good correlations of the color-difference formulae, such as  $\Delta E_{ab}^*$ , and the PSNR-based metrics. Therefore, it could be interesting to promote these metrics to compute image difference when alterations in contrast, lightness, and saturation are performed on a global scale, since they are computationally cheap and easy to implement.

### Gamut mapped images, Dugay et al.

In this dataset, 20 original images were gamut mapped with five different algorithms [22, 23]. The images were evaluated by 20 observers in a pair comparison experiment.

We see from the results in Figure 4 that most of the metrics fail utterly in evaluating perceived difference. This is probably because in gamut mapping multiple attributes are altered, therefore the objective assessment is very complex and the observers may judge the images differently [22, 23]. Previous research has also shown that IQ metrics have problems when multiple distortions occur simultaneously, as in gamut mapping [38, 39]. This is not the case for TID2008 and some of the other databases evaluated here, since usually only one attribute changes in the images at the time.

Metric / Database	Pedersen	Caracciolo	IVC	Ajagamelle	Dugay
$S_{DOG} - CIELAB1$	0.201	0.232	0.288	0.407	0.047
$S_{DOG} - CIELAB2$	0.200	0.196	0.268	0.483	-0.003
$S_{DOG} - CIELAB3$	0.193	0.170	0.262	0.532	-0.012
$S_{DOG} - CIELAB4$	0.197	0.178	0.269	0.504	-0.009
$S_{DOG} - DEE1$	0.142	0.052	0.425	0.551	0.039
$S_{DOG} - DEE2$	0.209	0.094	0.579	0.478	0.023
$S_{DOG} - DEE3$	0.201	0.146	0.655	0.474	0.020
$S_{DOG} - DEE4$	0.236	0.113	0.551	0.594	0.022
$S_{DOG} - DEE5$	0.079	0.010	0.285	0.413	0.027
S-CIELAB	0.798	0.242	0.705	0.675	-0.067
$\Delta E_{ab}^*$	0.764	0.156	0.539	0.751	-0.025
$\Delta E_E$	0.183	-0.041	0.023	0.597	-0.003
S-CIELAB <sub>JOHNSON</sub>	0.778	0.199	0.485	0.584	0.016
S-DEE	0.179	0.080	0.295	0.402	-0.002
Hue Angle	0.805	0.065	0.345	0.625	0.006
SHAME	0.802	0.171	0.662	0.499	0.042
SHAMEII	0.827	0.126	0.458	0.622	-0.100
UIQ	0.446	0.050	0.819	0.634	0.313
SSIM	0.217	0.294	0.705	0.635	0.159
SSIM-IPT	0.302	0.247	0.706	0.658	0.005
$Q_{COLOR}$	0.277	0.154	0.613	0.620	0.191
VSNR	X	0.043	0.782	X	0.087
VIF	0.393	0.214	0.880	0.530	0.314
HVS REAL	0.338	0.084	0.298	0.480	-0.025
PSNR-HVS	0.633	0.249	0.730	0.786	0.074
PSNR-HVS-M	0.626	0.311	0.734	0.785	0.073
ABF	0.805	0.163	0.060	0.734	0.084
$\Delta E_{94}^*$	0.650	0.130	0.369	0.743	0.016
$\Delta E_{00}^*$	0.626	0.134	0.383	0.739	0.034
MSE	0.666	0.264	0.510	0.614	0.086
RMS	0.750	0.267	0.605	0.746	0.063
PSNR	0.656	0.312	0.671	0.723	0.045
Structural Content	0.058	-0.255	0.234	-0.702	-0.017
Average Difference	-0.015	-0.207	-0.008	-0.733	-0.025
N-Cross-Correlation	0.004	0.082	-0.374	0.682	0.003
Correlation Quality	0.001	0.001	-0.060	0.165	0.004
Maximum Difference	0.108	0.451	0.653	-0.534	0.081
Image Fidelity	0.454	0.213	0.500	0.543	0.042

Table 4: Pearson correlation coefficients for the selected image databases. The results are highlighted for the metrics showing the best performance. "X" indicates that the result is not available for the metric.

## Performance of the IQMs

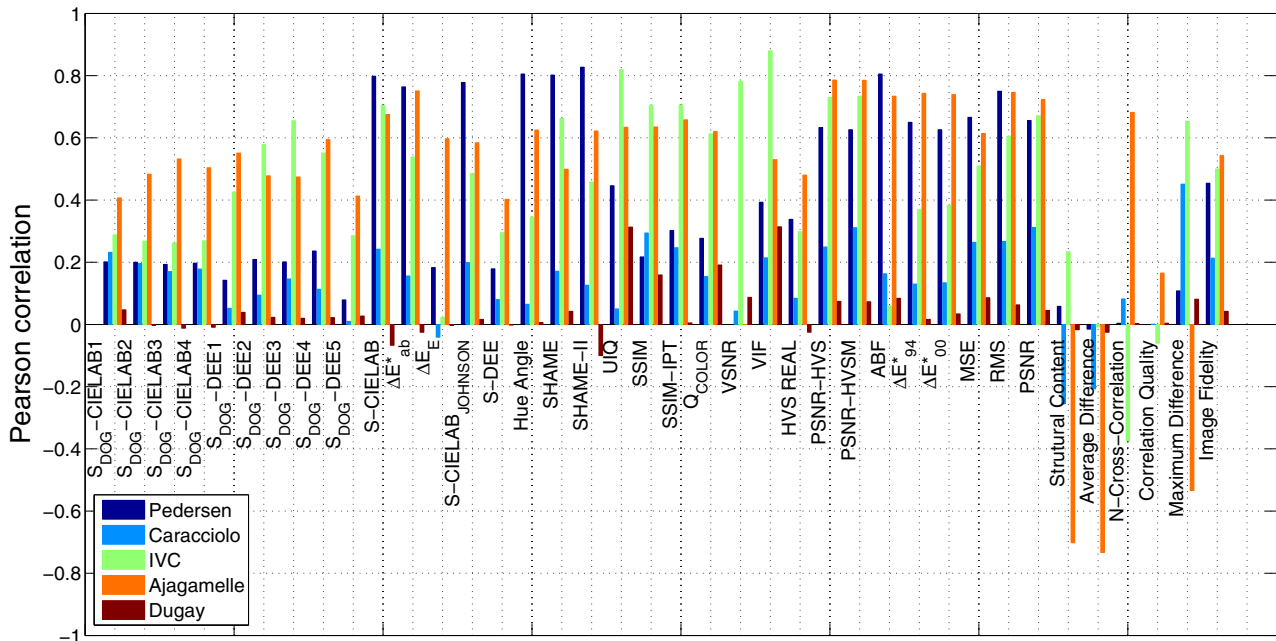


Figure 4: Performance of the IQ metrics across the databases. The figure shows the Pearson correlation coefficients for five different sets of test images. We found that  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$  have better correlations for certain artifacts or distortions, but do not outperform state-of-the-art IQ metrics.

## Conclusion and Perspectives

In this study, we proposed and assessed two new ID metrics named  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$ . It was shown that these metrics have better performance for certain attributes or distortions. We thought that the alternative ideas behind the DOG model would make them more efficient ID metrics. Nevertheless, the results indicate that none of them outperforms common state-of-the-art IQ metrics. Regardless of the color-difference formula that is chosen, the metrics do not provide high performance when extensively tested. We conclude that  $S_{DOG} - CIELAB$  and  $S_{DOG} - DEE$  are not yet effective measures of perceived difference. We did not achieve our objective of developing a reliable metric. Even so, an important stride has been taken towards a comprehensive ID evaluation.

In the future, we will seek to improve these trial models into more accurate and consistent metrics. An interesting line of inquiry is to use the DOG as a pre-filter in conjunction with a more classical spatial filtering module, that uses a CSF for example.

## References

- [1] M. Pedersen and J.Y. Hardeberg. Survey of full-reference image quality metrics. Høgskolen i Gjøviks rapportserie 5, The Norwegian Color Research Laboratory (Gjøvik University College), June 2009.
- [2] G.M. Johnson and M.D. Fairchild. From color image difference models to image quality metrics. In *International Congress on Imaging Science*, pages 326–327, Japan, 2002.
- [3] X. Zhang and B.A. Wandell. A spatial extension of CIELAB for digital color image reproduction. *Society for Information Display*, 5:61–63, 1997.
- [4] Y. Tadmor and D.J. Tolhurst. Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. *Vision Research*, 40:3145–3157, 2000.
- [5] A. Rizzi, G. Simone, and R. Cordone. A modified algorithm for perceived contrast in digital images. In *CGIV 2008 - Fourth European Conference on Color in Graphics, Imaging and Vision*, pages 249–252, Terrassa, Spain, Jun 2008. IS&T.
- [6] M. Pedersen and J. Y. Hardeberg. A new spatial hue angle metric for perceptual image difference. In Alain Trémeau, Raimondo Schettini, and Shoji Tominaga, editors, *Second International Workshop Computational Color Imaging (CCIW09)*, volume 5646 of *Lecture Notes in Computer Science*, pages 81–90, Saint-Etienne, France, Mar 2009. Springer.
- [7] G.M. Johnson and M.D. Fairchild. Darwinism of color image difference models. In *Proc. of IS&T/SID 9th Color Imaging Conference*, pages 108–112, 2001.
- [8] G. Simone, C. Oleari, and I. Farup. Performance of the Euclidean color-difference formula in log-compressed OSA-UCS space applied to modified-image-difference metrics. In *11th Congress of the International Colour Association (AIC)*, Sydney, Australia, Oct 2009.
- [9] C. Oleari, M. Melgosa, and R. Huertas. Euclidean color-difference formula for small-medium color differences in log-compressed OSA-UCS space. *Journal of the Optical Society of America A*, 26(1):121–134, 2009.
- [10] C. Oleari. Color opponencies in the system of the uniform color scales of the Optical Society of America. *Journal of the Optical Society of America A*, 21:677–682, 2004.
- [11] C. Oleari. Hypotheses for chromatic opponency functions and their performance on classical psychophysical data. *Color Research and Application*, 30(1):31–41, 2005.
- [12] M. Pedersen, N. Bonnier, J. Y. Hardeberg, and F. Albreghsen. Attributes of image quality for color prints. *Journal of Electronic Imaging*, 19(1):011016–1 – 011016–13, Jan 2010.
- [13] G. Simone, M. Pedersen, and J. Y. Hardeberg. Measuring perceptual contrast in digital images. *Journal of Visual Communication and Image Representation*, 2010, Submitted.

- [14] A. Rizzi, T. Algeri, G. Medeghini, and D. Marini. A proposal for contrast measure in digital images. In *CGIV 2004 – Second European Conference on Color in Graphics, Imaging and Vision*, pages 187–192, Aachen, Germany, April 2004. IS&T.
- [15] G. Simone, M. Pedersen, J.Y. Hardeberg, and A. Rizzi. Measuring perceptual contrast in a multilevel framework. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging XIV*, volume 7240, pages 72400Q.1–72400Q.8, San Jose, USA, Jan 2009. SPIE.
- [16] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - A database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, Vol. 10:30–45, 2009. <http://www.ponomarenko.info/tid2008.htm>.
- [17] P. Le Callet and F. Autrusseau. Subjective quality assessment IRCCyN/IVC database, 2005. <http://www.irccyn.ec-nantes.fr/ivcdb/>.
- [18] M. Pedersen. Importance of region-of-interest on image difference metrics. Master's thesis, Gjøvik University College, 2007.
- [19] M. Pedersen, J.Y. Hardeberg, and P. Nussbaum. Using gaze information to improve image difference metrics. In B. Rogowitz and T. Pappas, editors, *Human Vision and Electronic Imaging VIII (HVEI-08)*, volume 6806 of *SPIE proceedings*, pages 680611.1–680611.12, San Jose, USA, Jan 2008. SPIE.
- [20] V. Caracciolo. Just noticeable distortion evaluation in color images. Master's thesis, Gjøvik University College and Roma Tre University, 2009.
- [21] S.A. Ajagamelle. Analysis of the difference of Gaussians model in perceptual image difference metrics. Master's thesis, Gjøvik University College and Grenoble Institute of Technology, 2009.
- [22] F. Dugay. Perceptual evaluation of colour gamut mapping algorithms. Master thesis, Gjøvik University College and Grenoble Institute of Technology, 2007.
- [23] F. Dugay, I. Farup, and J.Y. Hardeberg. Perceptual evaluation of color gamut mapping algorithms. *Color Research & Application*, 33(6):470–476, Dec 2008.
- [24] G. Hong and M.R. Luo. Perceptually based colour difference for complex images. In R. Chung and A. Rodrigues, editors, *9th Congress of the International Colour Association*, volume 4421 of *Proceedings of SPIE*, pages 618–621, 2002.
- [25] G. Hong and M.R. Luo. New algorithm for calculating perceived colour difference of images. *Imaging Science Journal*, 54(2):86–91, 2006.
- [26] Z. Wang and A.C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9:81–84, 2002.
- [27] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [28] N. Bonnier, F. Schmitt, H. Brettel, and S. Berche. Evaluation of spatial gamut mapping algorithms. In *14th Color Imaging Conference*, volume 14, pages 56–61. IS&T/SID, Nov 2006.
- [29] A. Toet and M.P. Lucassen. A new universal colour image fidelity metric. *Displays*, 24:197–204, 2003.
- [30] D.M. Chandler and S.S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, 2007.
- [31] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.
- [32] T. Frese, C.A. Bouman, and J.P. Allebach. A methodology for designing image similarity metrics based on human visual system models. Technical Report TR-ECE 97-2, Purdue University, West Lafayette, IN, USA, 1997.
- [33] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli. Two new full-reference quality metrics based on HVS. In *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, 4 p., Scottsdale, Arizona, USA, 2006.
- [34] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin. On between-coefficient contrast masking of DCT basis functions. In *CD-ROM Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-07*, 4 p., Scottsdale, Arizona, USA, 25–26 January 2007.
- [35] Z. Wang and J.Y. Hardeberg. An adaptive bilateral filter for predicting color image difference. In *17th Color Imaging Conference*, pages 27–31, Albuquerque, NM, USA, Nov 2009.
- [36] A.M. Eskicioglu and P.S. Fisher. A survey of quality measures for gray scale image compression. In *1993 Space and Earth Science Data Compression Workshop*, pages 49–61, Snowbird, Utah, Apr. 2 1993. NASA Conference Publication 3191.
- [37] A.M. Eskicioglu, P.S. Fisher, and S. Chen. Image quality measures and their performance. In *IEEE Transactions on Communications*, volume 43, pages 2959–2965, Dec 1995.
- [38] J.Y. Hardeberg, E. Bando, and M. Pedersen. Evaluating colour image difference metrics for gamut-mapped images. *Coloration Technology*, 124(4):243–253, Aug 2008.
- [39] N. Bonnier, F. Schmitt, H. Brettel, and S. Berche. Evaluation of spatial gamut mapping algorithms. In *14th Color Imaging Conference*, volume 14, pages 56–61. IS&T/SID, Nov 2006.

## Author Biography

**Sebastien Ajagamelle** received the BSc degree in Engineering Sciences from Grenoble Institute of Technology, France and a MSc in Print Media in 2009 from Grenoble INP-Pagora, a department of Grenoble Institute of Technology. He is currently working for FIROPA Printing Group, France.

**Marius Pedersen** received his BSc in Computer Engineering in 2006, and MiT in Media Technology in 2007, both from Gjøvik University College, Norway. He is currently pursuing a PhD in Color Imaging, under the supervision of Pr. Hardeberg and Pr. Albrechtsen, sponsored by Océ. He is also a member of the Norwegian Color Research Laboratory at Gjøvik University College. His work is centered on image quality metrics for color prints.

**Gabriele Simone** received his BiT in 2005, and his MSIT in 2007 both at University of Milan - Department of Information Technology, Italy. He is currently pursuing a PhD at Gjøvik University College, Norway. His main research topic is contrast measure, image difference metrics, and tone mapping algorithms in HDR images.