# Evaluation of a Difference of Gaussians Based Image Difference Metric in Relation to Perceived Compression Artifacts

Gabriele Simone[1], Valentina Caracciolo[1,2], Marius Pedersen[1], and Faouzi Alaya Cheikh[1]

[1] Gjøvik University College, Gjøvik, Norway
[2] University of Roma Tre, Rome, Italy

**Abstract.** In this paper we investigate if the Difference of Gaussians model is able to predict observers perceived difference in relation to compression artifacts. A new image difference metric for specifically designed for compression artifacts is proposed. In order to evaluate this new metric a psychophysical experiment is carried out, where a dataset of 80 compressed JPEG and JPEG2000 images were generated from 10 different scenes. The results of the psychophysical experiment with 18 observers and the quality scores obtained from a large number of image difference metrics are presented.

Furthermore, a quantitative study based on a number of image difference metrics and five additional databases is performed in order to reveal the potential of the proposed metric. The analyses show that the proposed metric and most of the tested ones do not correlate well with the subjective test results, and thus the increased complexity of the recent metrics is not justified.

## 1 Introduction

We are witnessing a rapid growth in the number of media-rich documents that are created, transmitted, and consumed in our everyday activities. Furthermore, multimedia applications are also developing fast and becoming able to manage rather large images (e.g. image archiving, network image transmission, document imaging, digital photography, medical imaging, remote sensing, ...). Therefore, the demand for efficient and versatile image compression is more pressing than it has ever been before. Additionally, the improved media reproduction hardware has put a strain on the acceptable level of the quality of the processed media. During acquisition, communication, and consumption digital images are subjected to a wide variety of distortions, such as errors, noise and in particular compression artifacts. To ensure a certain quality level of multimedia applications/services one needs automatic means to evaluate the quality of digital images. Typically to identify the best reproduction among a number of variants of one reproduction algorithm (e.g. JPEG or JPEG2000 for compression), a psychophysical experiment has to be carried out. This results in a scale with

G. Bebis et al. (Eds.): ISVC 2010, Part II, LNCS 6454, pp. 491–500, 2010.

the perceived visual difference of the reproductions from the original image. Psychophysical experiments are very good in estimating the perceived distortion/quality of the media, they are, however, both time and resource demanding. Furthermore, they are not practical to use when automatic quality evaluation is needed, for example in computer vision based systems. This is why objective image difference metrics have been introduced. These are in general very simple mathematical formulae that give an objective measure of the quality of a reproduction, e.g PSNR, RMS, and MSE. They are however not sensitive to what the Human Visual System (HVS) is sensitive to.

Image difference metrics are based on a number of different ideas, even so, these metrics usually follow a general framework. In the most common framework the image and its reproduction are 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 of the image by a local neighbor to more complex methods, such as using Contrast Sensitivity Function(CSF). Finally the difference between the two images is calculated, in general, using a color difference formula. This accounts for the outstanding number of image difference metrics that have been developed so far [1].

The rest of this paper will be organized as follows: Section 2 provides an insight into the state-of-the-art of image difference metrics. Section 3 describes the proposed metric, while Section 4 describes the psychophysical experiment. Section 5 presents the experimental results and discusses how the results from the image difference metrics reflects their perceptual quality estimation. Finally, in section 6 conclusions are drawn.

## 2 State of the Art of Image Difference Metrics

In 1976, CIE published the CIELAB color space as a uniform color space, in which the difference between two colors $\Delta E^*_{ab}$ is represented by their Euclidean distance. The CIELAB metric has been used as a tool for measuring perceptual differences between uniform patches of colors. Although non-appropriate, the CIELAB $\Delta E^*_{ab}$ has been used for measuring the color difference between images by computing the color difference of all the pixels and averaging. The unsatisfactory uniformity of CIELAB space induced researchers to produce other color-difference data and search for better color-difference formulae. The last CIE formula for small-medium color differences is the $\Delta E^*_{00}$ one, termed CIEDE2000 and based on a wider set of empirical data [2].

In 1997, Zhang and Wandell proposed a spatial extension to the CIELAB color-difference formula, termed $S-CIELAB$ [3], that provides both a spatial filtering to simulate the blurring of the HVS and a consistency with the $CIELAB$ for large uniform areas [4]. In 2001 Johnson and Fairchild followed a similar approach [5], where the spatial filter is implemented in the frequency domain, obtaining a more precise control of the CSFs.

The Hue Angle Algorithm proposed by Hong and Luo in 2002 [6] is based on the known fact that systematic errors over the entire image is quite noticeable and unacceptable. A histogram based of the hue angle is computed, and sorted in ascending order so that weights can be applied to four different quartiles of the histogram. The overall color difference is calculated by multiplying the weighted hue angle for every pixel. The Spatial Hue Angle Metric ($SHAME$), proposed by Pedersen and Hardeberg [1], can be considered as the combination of the original $S - CIELAB$ and the hue angle algorithm, thus taking into account the spatial properties of the HVS. $SHAME - II$ is a variation of $SHAME - I$ that applies the filtering proposed by Johnson and Fairchild before applying the hue angle metric.

Very recently, in 2009, a Euclidean color-difference formula for small-medium color differences in log-compressed OSA-UCS space, termed $\Delta E_E$, has been published by Oleari et al. [7]. This formula is statistically equivalent to CIEDE2000 in the prediction of many available empirical datasets, but with greater simplicity and clear relationship with visual processing [8]. In 2009 Simone et al. [9] proposed and tested a new metric, named $S{-}\Delta E_E$, which works as the $S{-}CIELAB$ from Johnson and Fairchild, but the $\Delta E_{ab}$ is substituted with the $\Delta E_E$. Following, Ajagamelle [10] applied a novel technique, developing $S_{DOG} - CIELAB$ which uses the Difference of Gaussians (DOG) model as a basis for the spatial filtering with the $\Delta E^*_{ab}$ as a color difference formula, and $S_{DOG} - DEE$ which uses the DOG model and $\Delta E_E$ as color difference formula. The $\Delta E_E$ and the derivate image difference metrics have been extensively tested by Simone et al. [9] and Ajagamelle et al. [11].

Universal Image Quality (UIQ) proposed by Wang and Bovik [12] models image distortions as the combination of three elements: loss of correlation, luminance distortion and loss of contrast. Structural Similarity (SSIM) from Wang et al. [13] introduces the possibility of choosing the importance exponent for each of the three factors.

Many other image difference metrics using different approaches are available in the literature. For an extensive and detailed description of these metrics we refer the reader to the survey from Pedersen and Hardeberg [1].

## 3 The New Metric

Recent studies have shown that contrast is an important image attribute that falls under the umbrella of image quality [14]. In 2000 Tadmor and Tolhurst [15] 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:

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

where $x$ and $y$ indicate the row and the column of the pixel $(x, y)$, and $r_c$ is the radius of the Gaussian. The surround component is represented by a Gaussian curve as well, with a larger radius $r_s$:

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

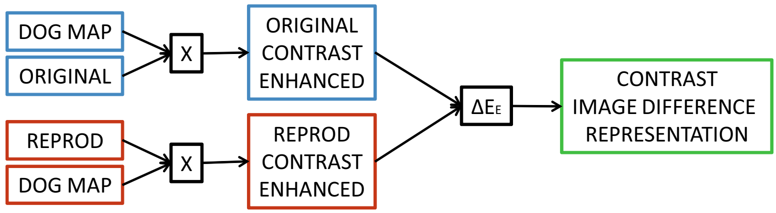


**Fig. 1.** Workflow of $M_{DOG} - DEE$

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:

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

while the output of the surround component is:

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

The following criterion for the measure of contrast was proposed, where the response gain is set by the local mean luminance:

$$C(x, y) = (R_c(x, y) - R_s(x, y))/(R_c(x, y) + R_s(x, y)).$$

The DOG model has revealed to be beneficial in the identification of edges and blocks, and it has been extensively tested as contrast measure by Simone et al. [16] and as image difference metric by Ajagamelle et al. [11].

In our novel approach the DOG model is used as weighting map for each pixel before applying a color difference formula, in order to give more importance to those regions where edges and blocks can appear due to compression artifacts. As color difference formula we have selected the $\Delta E_E$ from Oleari et al. [7]. The workflow of this metric, that we call $M_{DOG} - DEE$ is shown in Figure 1.

We selected several combinations of $r_c$ and $r_s$ in order to see whether a particular configuration would result in better adequacy with the subjective evaluation. We have chosen the following combinations as suggested in [16]: $r_c = 1$ and $r_s = 2$ (C1); $r_c = 1$ and $r_s = 3$ (C2); $r_c = 2$ and $r_s = 3$ (C3); $r_c = 3$ and $r_s = 4$ (C4).

## 4 Psychophysical Experiment

### 4.1 Creating the Dataset

A total of 10 images have been chosen, covering a wide range of distortions and a wide range of scenes, for the evaluation of several image difference metrics. The test database has been selected according to the quality attributes that are


**Fig. 2.** Scenes used for creating the dataset

being evaluated following the recommendations of Field [17] and the CIE [18]. The chosen images are shown in Figure 2. Three images (Figure 2(c), Figure 2(d) and Figure 2(i)) were captured and provided by one of the authors; Figure 2(e), Figure 2(g) and Figure 2(j) were selected from TID2008 database [19]; Figure 2(h) was selected from a standard natural image set provided by the CIE [20]; Figure 2(a) was selected from High Dynamic Range Imaging - Acquisition, Display and Image-Based Lighting [21]; Figure 2(b) was selected from Alan Gersho's lab at U.C. Santa Barbara; Figure 2(f) was selected from Le Callet database [22].

Four different levels of compression were used for JPEG and JPEG200, resulting in eight compressed images for each scene and a total number of 80 test images. The compression rate in bit per pixel (*bpp*) was calculated as:

$$bpp = (file\ size\ in\ bytes \times 8 / image\ size\ in\ pixels).$$

For the experiment the compression rates were chosen such that JPEG and JPEG2000 have similar values of the *bpp*. Table 1 reports the *bpp* used to compress each scene. All the compressed images were generated with the quality

**Table 1.** Selected *bpps* for each image

| Figure | *bpp* JPEG and JPEG2000 | | | |
|---|---|---|---|---|
| 2(a) | 0.8783 | 0.9331 | 1.0074 | 1.1237 |
| 2(b) | 0.9614 | 1.0482 | 1.1003 | 1.2569 |
| 2(c) | 0.8293 | 0.9080 | 0.9869 | 1.0977 |
| 2(d) | 0.6611 | 0.7226 | 0.7844 | 0.8726 |
| 2(e) | 0.9817 | 1.0500 | 1.1445 | 1.2916 |
| 2(f) | 1.4482 | 1.5910 | 1.7057 | 1.9366 |
| 2(g) | 0.6230 | 0.6891 | 0.7372 | 0.8169 |
| 2(h) | 1.0353 | 1.1338 | 1.2351 | 1.3765 |
| 2(i) | 0.4148 | 0.4474 | 0.4875 | 0.5420 |
| 2(j) | 0.7518 | 0.8198 | 0.8774 | 0.9927 |

factor in the range 0-100 for JPEG and JPEG2000 with the associated *bpp* and a pre-test was carried out to choose the range of the just noticeable distortion (JND).

## 4.2 Experimental Setup

The experiment was conducted using the category judgement method [23]. The observer was instructed to judge an image according to quality, where the quality of the image is assigned to one of seven categories. All pairs of images were presented to the 18 recruited observers with the original one on the left and one compressed with random *bpp* and random type of compression (JPEG or JPEG2000) on the right side of the display. Each pair of images were displayed on an Eizo ColorEdge CG241W digital LCD display. The monitor was calibrated and profiled using GretagMacbeth, Eye-One Match 3. The settings on the monitor were sRGB with 40% brightness and a resolution of 1600 × 1200 pixels. The experiment took place in a windowless room with neutral grey walls, ceiling and floor. The ceiling lights in the room was set to provide a level of ambient illumination around 32 lux, which is below the upper threshold recommended by the CIE (64 lux) [24]. The white point was set to the D65 white point and the gamma is set to a value of 2.2. The display was placed at a viewing distance of 70 cm.

From the pair comparison experiment z-scores were calculated. In Figure 3 the z-scores with confidence intervals for all images are displayed. The results that we obtained are in agreement with what we expected, we have obtained an ascending order, from the image with the lowest *bpp* to the image with the highest *bpp* for JPEG and the same for JPEG2000. As the JPEG2000 bottom value is higher than JPEG we can have higher compression with small differences.

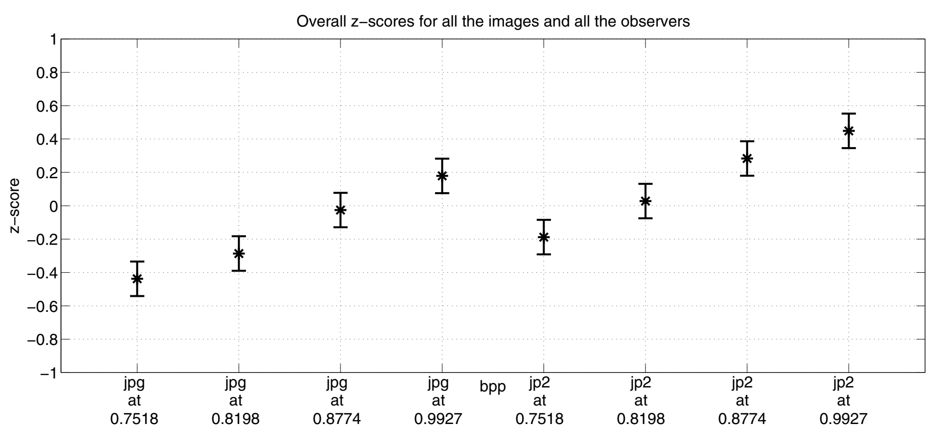


**Fig. 3.** Z-scores of all images

For this we can claim that JPEG2000 seems to perform better than JPEG (in agreement with the theory) [25]. The range between the highest and lowest values for JPEG2000 is larger than in the range in the JPEG and because of this it seems we have more perceived difference among the compression levels. The range between the highest and the lowest values in Figure 3 is quite narrow, in accordance with the fact that we were interested in evaluating the JND.

## 5 Experimental Results

In order to reveal potential differences between the methods and how $M_{DOG} - DEE$ performs on this particular dataset, two types of correlation were computed: the Pearson product-moment Correlation Coefficient (CC), which assumes that the variables are ordinal, and finds the linear relationship between them; the Spearman rank CC, which is a non-parametric measure of correlation that uses the ranks as basis instead of the actual values, thus the relationship between the variables is described without making any assumptions about the frequency distribution.

**Table 2.** Pearson CC for the tested metrics. In cyan the most performant metrics, while in red the least performant ones.

| Metric—Database | Proposed | Pedersen | IVC | Ajagamelle | Fabienne | TID2008 |
|---|---|---|---|---|---|---|
| $M_{DOG} - DEE$ (C1) | 0.078 | 0.178 | -0.010 | 0.709 | 0.034 | -0.201 |
| $M_{DOG} - DEE$ (C2) | 0.086 | 0.188 | 0.011 | 0.698 | 0.048 | -0.242 |
| $M_{DOG} - DEE$ (C3) | 0.081 | 0.188 | 0.025 | 0.699 | 0.029 | -0.305 |
| $M_{DOG} - DEE$ (C4) | 0.087 | 0.183 | 0.035 | 0.692 | 0.026 | -0.315 |
| Maximum Difference | 0.451 | 0.108 | 0.653 | -0.534 | 0.081 | 0.189 |
| UIQ | 0.050 | 0.446 | 0.819 | 0.634 | 0.313 | 0.622 |
| SSIM | 0.294 | 0.217 | 0.705 | 0.635 | 0.159 | 0.550 |
| $\Delta E^*_{ab}$ | 0.156 | 0.764 | 0.539 | 0.751 | -0.025 | 0.232 |
| S-CIELAB | 0.242 | 0.798 | 0.705 | 0.675 | -0.067 | 0.433 |
| $S-CIELAB_{JOHNSON}$ | 0.199 | 0.778 | 0.485 | 0.584 | 0.016 | 0.318 |
| $S_{DOG} - CIELAB$ | 0.232 | 0.201 | 0.288 | 0.407 | 0.047 | 0.385 |
| Hue Angle | 0.065 | 0.805 | 0.345 | 0.625 | 0.006 | 0.262 |
| SHAME | 0.171 | 0.802 | 0.662 | 0.499 | 0.042 | 0.300 |
| SHAMEII | 0.126 | 0.827 | 0.458 | 0.622 | -0.100 | 0.408 |
| $\Delta E_E$ | -0.041 | 0.183 | 0.023 | 0.597 | -0.003 | 0.273 |
| $S - DEE$ | 0.080 | 0.179 | 0.295 | 0.402 | -0.002 | 0.294 |
| $S_{DOG} - DEE$ | 0.146 | 0.201 | 0.655 | 0.474 | 0.020 | 0.328 |
| $\Delta E^*_{00}$ | 0.134 | 0.626 | 0.383 | 0.739 | 0.034 | 0.086 |
| RMS | 0.267 | 0.750 | 0.605 | 0.746 | 0.063 | 0.536 |
| MSE | 0.264 | 0.666 | 0.510 | 0.614 | 0.086 | 0.535 |
| Structural Content | -0.255 | 0.058 | 0.234 | -0.702 | -0.017 | 0.025 |
| Average Difference | -0.207 | -0.015 | -0.008 | -0.733 | -0.025 | 0.153 |
| PSNR | 0.312 | 0.656 | 0.671 | 0.723 | 0.045 | 0.508 |

Table 2 shows the Pearson CC of the proposed metric and other state of the art metrics in addition to color difference formulae and several numerical objective quality measure as MSE, RMS, PSNR, structural content, average difference, and maximum difference [26]. Due to page limitations we will present only a selection of the calculated results.

$M_{DOG} - DEE$ shows a really poor correlation, also using different combinations of $r_c$ and $r_s$ indicating that the metric is not able to predict observers perceived difference in relation to compression artifacts. Maximum Difference and PSNR are the most performant metrics while Structural Content, Average Difference and $\Delta E_E$ are the least performant, showing a negative correlation. It is interesting to see that all the metrics using $\Delta E^*_{ab}$ perform better than the ones using $\Delta E_E$, and RMS and MSE, perform better than most of the metrics. The overall conclusion is that none of the tested metrics show a good correlation, indicating an inefficiency in predicting perceived compression on this dataset. Spearman CC gives very similar results.

In order to investigate the performance of $M_{DOG} - DEE$ extensively and to see if it is suitable for other image quality purposes, we have tested it on five other datasets:

- Luminance changed images: this database from Pedersen [27] includes four original images reproduced with different changes in lightness. Each scene has been altered in four ways globally and four ways locally.
- IVC database: the IVC database from Le Callet et al. [22] contains blurred images and images distorted by three types of lossy compression techniques (JPEG, JPEG2000, and Locally Adaptative Resolution).
- Images altered in contrast, lightness, and saturation: this database from Ajagamelle contains a total of 10 original images covering a wide range of characteristics and scenes [10]. The images were modified on a global scale with separate and simultaneous variations of contrast, lightness, and saturation.
- Gamut mapped images: this database from Dugay et al. [28] is composed of 20 original images, which were gamut mapped with five different algorithms.
- TID2008: this database from Ponomarenko et al. [19] contains a total of 1700 images, with 25 reference images and 17 types of distortions over 4 distortion levels.

$M_{DOG} - DEE$ shows good correlation only on the database of images altered in contrast, lightness, and saturation from Ajagamelle [10], but it does not outperform other state of the art metrics and it turns out to be the least performant on the TID2008 database. Furthermore it is interesting to notice that none of the tested metrics show a good correlation in the database of gamut mapped images from Dugay et al. [28]. For all the databases Sperman CC gives very similar results. In conclusion the results from this analysis supports the findings from the first dataset, where using the DOG model as weighting map does not improve the performance of state of the art metrics in predicting perceived compression distortions and image quality.

## 6 Conclusions and Perspectives

In this paper the Difference of Gaussians model has been investigated with the purpose to see if it is able of predicting observers perceived difference in relation to compression artifacts. For that purpose a new image difference metric has been developed and a psychophysical experiment was conducted. A dataset of 80 compressed images were generated from 10 different scenes, based on different levels of compression via JPEG and JPEG2000. From psychophysical experiment it is easy to see that as expected JPEG2000 performs better than JPEG, having higher compression with small differences in accordance to observers perceived distortion. The metric developed and most of the tested state of the art image difference metrics show a poor correlation with the viewers perceptual quality ratings, indicating an inefficiency in predicting perceived compression distortions on this dataset. Furthermore, from extensive tests on five other databases, it can clearly be seen that the increased computational complexity of the recent image difference metrics are not proportional with the obtained performance improvement over simpler metrics.

## References


1. Pedersen, M., Hardeberg, J.Y.: Survey of full-reference image quality metrics. Høgskolen i Gjøviks rapportserie, vol. 5. Gjøvik University College, The Norwegian Color Research Laboratory, Gjøvik, Norway (2009)
2. Luo, M., Cui, G., Rigg, B.: The development of the CIE 2000 colour-difference formula: CIEDE2000. Color Research and Application 26, 340–350 (2001)
3. Zhang, X.M., Farrell, J., Wandell, B.: Application of a spatial extension to cielab. In: IS&T/SPIE Electronic Imaging 1997, vol. 3025, pp. 154–157 (1997)
4. Johnson, G.M.: Measuring images: differences, quality and apperance. PhD thesis, Rochester Institute of Technology (2003)
5. Johnson, G.M., Fairchild, M.D.: Darwinism of color image difference models. In: IS&T/SID 9th Color Imaging Conference, Scottsdale, AZ, USA, pp. 108–112 (2001)
6. Hong, G., Luo, M.R.: New algorithm for calculating perceived colour difference of images. Imaging Science Journal 54, 86–91 (2006)
7. Oleari, C., Melgosa, M., Huertas, R.: Euclidean color-difference formula for small-medium color differences in log-compressed osa-ucs space. Journal of the Optical Society of America 26, 121–134 (2009)
8. Huertas, R., Melgosa, M., Oleari, C.: Performance of a color-difference formula based on OSA-UCS space using small-medium color differences. Journal of the Optical Society of America 23, 2077–2084 (2006)
9. Simone, G., Oleari, C., Farup, I.: 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, p. 81 (2009)
10. Ajagamelle, S.: Analysis of the difference of gaussians model in perceptual image difference metrics. Master's thesis, Gjøvik University College and Grenoble Institute of Technology (2009)
11. Ajagamelle, S.A., Pedersen, M., Simone, G.: Analysis of the difference of gaussians model in image difference metrics. In: 5th European Conference on Colour in Graphics, Imaging, and Vision (CGIV), Joensuu, Finland, pp. 489–496 (2010)

12. Wang, Z., Bovik, A.: A universal image quality index. IEEE Signal Processing Letters 9, 81–84 (2002)
13. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP 13, 600–612 (2004)
14. Pedersen, M., Bonnier, N., Hardeberg, J.Y., Albregtsen, F.: Attributes of image quality for color prints. Journal of Electronic Imaging 19, 011016-1– 011016-13 (2010)
15. Tadmor, Y., Tolhurst, D.: Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. Vision Research 40, 3145–3157 (2000)
16. Simone, G., Pedersen, M., Hardeberg, J.Y.: Measuring perceptual contrast in digital images. Journal of Visual Communication and Image Representation (2010) (under review)
17. Field, G.G.: Test image design guidelines for color quality evaluation. In: IS&T/SID 7th Color Imaging Conference, Scottsdale, AZ, USA, pp. 194–196 (1999)
18. CIE: Guidelines for the evaluation of gamut mapping algorithms. Technical Report, CIE TC8-08 (156:2004) ISBN: 3-901-906-26-6
19. Ponomarenko, N., Lukin, V., Egiazarian, K., Astola, J., Carli, M., Battisti, F.: Color image database for evaluation of image quality metrics, pp. 403–408 (2008), `http://www.ponomarenko.info/tid2008.htm`
20. ISO: Graphic techonology - prepress digital echange. Technical report, ISO 12640-2, 1 edn. (2004)
21. Reinhard, E., Ward, G., Pattanaik, S., Debevec, P.: High Dynamic Range Imaging - Acquisition, Display and Image-Based Lighting. Morgan Kaufmann Publisher, San Francisco (2005)
22. Le Callet, P.A.: Subjective quality assessment irccyn/ivc database 2005. In: IRCCyN (2005)
23. Engeldrum, P.G.: Psychometric Scaling. Imcotek Press, Winchester (2000)
24. Djik, J.: In search of an objective measure for the perceptual quality of printed images. PhD thesis, Technische Unisersitet Delft (2004)
25. Caracciolo, V.: Just noticeable distortion evaluation in color images. Master's thesis, Gjøvik University College and Roma Tre University (2009)
26. Eskicioglu, A., Fisher, P., Chen, S.: Image quality measures and their performance. IEEE Transactions on Communications 43, 2959–2965 (1995)
27. Pedersen, M.: Importance of region-of-interest on image difference metrics. Master thesis, Gjøvik University College (2007)
28. Dugay, F., Farup, I., Hardeberg, J.Y.: Perceptual evaluation of color gamut mapping algorithms. Color Research & Application 33, 470–476 (2008)