

A Multidistortion Database for Image Quality

Silvia Corchs^(✉) and Francesca Gasparini

Dipartimento di Informatica, Sistemistica e Comunicazione,
University of Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy
{corchs, gasparini}@disco.unimib.it

Abstract. In this paper we introduce a multidistortion database, where 10 pristine color images have been simultaneously distorted by two types of distortions: blur and JPEG and noise and JPEG. The two datasets consist of respectively 350 and 400 images, and have been subjectively evaluated within two psycho-physical experiments. We here also propose two no reference multidistortion metrics, one for each of the two datasets, as linear combinations of no reference single distortion ones. The optimized weights of the combinations are obtained using particle swarm optimization. The different combinations proposed show good performance when correlated with the subjective scores of the multidistortion database.

Keywords: Image quality assessment · Multidistortion database · No reference metrics · Blur · Noise · JPEG

1 Introduction

Image quality studies mainly focus on images corrupted by single distortions. However, consumer images suffer in general of more than one distortion simultaneously due to the different process that take place within their production flow (acquisition, compression, transmission, etc.), [1]. The vast majority of No Reference (NR) metrics have been developed to measure single distortions. In the last years, some NR metrics have also addressed multiple artifacts, most commonly blur and noise [2–5]. Also general purpose (or blind) NR metrics have been proposed that do not aim to detect specific types of distortion. These last methods approach the Image Quality Assessment (IQA) as a classification and regression problem in which the regressors/classifiers are trained using specific features obtained from natural-scene-statistics [6, 7]. Following Mittal et al. [6] it is also possible to individuate two subcategories of blind models: the Opinion-Aware (OA) models, that have to be trained on a database of human rated distorted images and associated subjective opinion scores, and the Opinion-Unaware (OU) ones. An overview of the different objective and subjective IQA methods can be found in the review articles [8, 9]. It is well known that any objective metric must be validated with respect to user judgments: subjective tests are at the base of objective quality metrics benchmarking and IQA databases

serve as ground-truth information for evaluating IQA algorithms. In general, the available databases contain images corrupted by only one of several possible distortions.

Considering multiply distorted images, Jayaraman et al. [10] have presented the LIVE-MD database. As most publicly available image quality databases, it has been created under highly controlled conditions by introducing graded simulated distortions onto high-quality images. Two scenarios are considered in LIVE-MD: images first blurred and then JPEG compressed (part 1), and images first blurred and then corrupted by white Gaussian noise (part 2). Two psycho-physical experiments have been conducted by the authors to collect the subjective data. Recently, Ghadiyaram and Bovik [11] presented the LIVE In the Wild Image Quality Challenge Database, that contains diverse authentic multi-distorted images. The images were collected without artificially introducing any distortion beyond those occurring during capture, processing, and storage by a users device. The authors conducted a very large-scale image quality assessment subjective implementing an online crowdsourcing system.

To compare objective and subjective results different performance measures are used. The Video Quality Experts Group (VQEG) [12] recommends three performance criteria for the metrics: prediction accuracy, prediction monotonicity and prediction consistency with respect to the subjective assessments. The prediction accuracy is quantified by the Pearson Correlation Coefficient (PCC) and the Root Mean Squared Error (RMSE). The Spearman Rank Order Correlation Coefficient (SROCC) measures the prediction monotonicity of a metric and the Outlier Ratio (OR) the prediction consistency. Before computing these correlation coefficients, it is customary to apply a nonlinear transformation to the predicted scores so as to bring the predictions on the same scale as the subjective scores in order to obtain a linear relationship between the predictions and the opinion scores. The VQEG suggests the use of logistic or polynomial functions. The parameters of these functions are chosen to minimize the MSE between the set of subjective values (of a particular database) and the corresponding set of transformed predicted values.

In this paper we present the Multiple Distorted IVL Database (MD-IVL). The database is composed of two parts: the MD-IVL-BJ that contains color images first blurred and then JPEG compressed, and the MD-IVL-NJ that contains color images corrupted by Gaussian noise and then JPEG compressed. Subjective studies were conducted on this database to obtain human judgments for the multiply distorted images and the corresponding psycho-visual data were collected.

We here propose to define NR-MD-metrics as linear combination of NR state-of-the-art metrics specifically designed for single distortions (blockiness, sharpness and noise). In particular we propose two different linear combinations, one for the MD-IVL-NJ dataset and the second for the MD-IVL-BJ one. The weighting coefficients of the linear combinations are obtained using a particle swarm optimization method [13]. We used half of the data to evaluate the coefficients and the remaining half to test the correlation performance of the proposed metrics for each of the two parts of the MD-IVL database. We also compare the

performance of the two linear combination proposals with two competitive blind models available in the literature: an OA model, BRISQUE [6] and an OU model, NIQE [7]. Moreover we test the linear combination proposed for the MD-IVL-BJ on part 1 of the LIVE-MD, as it is composed by images suffering the same type of distortions.

2 The MD-IVL Database

The MD-IVL database originates from 10 reference color images of the SD-IVL dataset [14]. The images of 886×591 pixels (15×10 cm at 150 dpi, typical printing parameters for natural photos) are chosen to sample different contents both in terms of low level features (frequencies, colors) and higher ones (face, buildings, close-up, outdoor, landscape). The corresponding thumbnails are shown in Fig. 1.



Fig. 1. The 10 reference images of the SD-IVL database.

Starting from these images we have generated:

- A database MD-IVL-BJ of 350 blur plus JPEG distorted images. These distorted images have been generated as follows: each of the reference images has been corrupted with seven levels of Gaussian blur corresponding to standard deviations of 0.001, 0.66, 1.33, 2, 2.66, 3.33 and 4. Each of the 70 blurred images has been further corrupted with four levels of JPEG compression corresponding to Q-factor values of 100, 50, 30, 20 and 10.
- A database MD-IVL-NJ of 400 noise plus JPEG distorted images. These distorted images have been generated as follows: each of the 10 reference images has been corrupted with ten levels of Gaussian noise corresponding to 1, 2, 3, 4, 5, 6, 8, 10, 12 and 14 gray levels of standard deviation on the luminance channel. Each of the 100 noisy images has been further processed by 4 different levels of JPEG compression, corresponding to Q factor values of 100, 50, 30, and 10.

As an example, we show in Fig. 2 the most distorted blur-JPEG and noise-JPEG images.



Fig. 2. (a) A reference image from SD-IVL database and its most multiply distorted versions for blur-JPEG, and noise-JPEG.

For collecting the subjective data, we have adopted a Single Stimulus method (SS) [15], where all the images are individually shown. The observers were asked to rate the images within a continuous scale from 0 (Worst quality) to 100 (Best quality). The experiments were performed following the recommendations in ITU [15].

The subjective study was conducted in two parts: each of the databases (MD-IVL-BJ and MD-IVL-NJ) has been assessed separately. Ideally, all images in a subjective QA study should be evaluated in only one session. However, it is recommended a maximum duration of 30 min for each experimental session [15]. Therefore, to assess the entire MD-IVL database we have conducted the experiments in several sessions using different sets of images. Each session consists of approximately 100 images (chosen so as to span all the distortion range) and has been evaluated by 12 observers.

Mean subjective scores were computed for each observer as follows. The raw, subjective score r_{ij} for the i -th subject ($i = 1, \dots, S$, with S = number of subjects) and j -th image I_j ($j = 1, \dots, N$, with N = number of dataset images) was converted into its corresponding Z-score as follows:

$$z_{ij} = \frac{r_{ij} - \bar{r}_i}{\sigma_i} \quad (1)$$

where \bar{r}_i and σ_i are the mean and the standard deviation of the subjective scores over all images assessed by the i -th subject.

Data were cleaned using a simple outlier detection algorithm. A score for an image was considered to be an outlier if it fell outside an interval of two standard deviations width about the mean score for that image across all subjects.

The remaining Z-scores, were then averaged across subjects to yield the Mean Opinion Scores y_j for each image j :

$$y_j = \frac{1}{S} \sum_{i=1}^S z_{ij} \quad (2)$$

The MD-IVL database is available at: <http://www.ivl.disco.unimib.it/activities/imagequality/>.

3 Objective Data

The subjective scores described in Sect. 2, collected in terms of MOS can be correlated with different NR metrics, using a logistic function.

Denoting by y_i the MOS value of the i -th image of the database ($i = 1, \dots, N$ with N the total number of distorted images) and by x_i the corresponding objective metric value, the logistic transformation reads:

$$f(x) = \frac{\alpha}{1 + \exp(\beta(x - \gamma))} + \delta \quad (3)$$

where the parameters α , β , γ and δ are chosen to minimize the mean square error between the subjective scores $\{y_i\}$ and the predicted ones $\{f(x_i)\}$.

3.1 NR Metrics for Single Distortion

Among the many distortion-specific NR metrics available in the literature, we have here chosen those that highly correlate with the corresponding subjective data in the case of blur, noise and JPEG artifacts. The metrics considered in this work are:

- M1: The bluriness metric by Marziliano et al. [16]. An edge detector is applied and for pixels corresponding to an edge location, the start and end positions of the edge are defined as the local extrema locations closest to the edge. The edge width is measured and identified as the local blur measure. Global blur obtained by averaging the local blur values over all edge locations.
- M2: metric Q by Zhu and Milanfar [5]. This measure is correlated with the noise level, sharpness and intensity contrast of the structured regions of an image. Its value drops if the variance of noise rises, and/or if the image content becomes blurry.
- M3: The JPEG-blockiness metric by Wu and Yuen [17], named Generalized Block-edge Impairment. It is the most well known metric in the spatial domain. It measures the blockiness separately in horizontal and vertical direction, after which the two directions are combined into a single quality value.
- M4: The JPEG-blockiness specific metric by Wang et al. [18]. It is formulated in the frequency domain and models the blocky image as a non-blocky image interfered with a pure blocky signal.
- M5: The noise metric by Immerkaer [19]. It estimates the standard deviation of Additive White Gaussian Noise from a single image using a Laplacian mask filtering approach.

All of these metrics show high performance in terms of PCC when correlated with well known databases of single distorted images like LIVE [20], CSIQ [21] and SD-IVL [14] among others.

3.2 NR Metric for the Multiply Distorted Images: Our Proposal

We here propose a MD-NR metric as a linear combination (LC) of single distortion NR metrics:

$$LC(I_j) = \sum_{k=1}^K a_k \times M_k(I_j) \quad (4)$$

where I_j is the j -th image of the dataset ($j = 1, \dots, N$), and M_k is the k -th single distortion NR metric among the K considered ones. In this paper we propose to set the weighting coefficients $\{a_k\}$ using a population based stochastic optimization technique, called Particle Swarm Optimization (PSO) [22, 23] that optimally fits the subjective scores of a proper MD database.

Recalling that one of the criteria recommended by the VQEG [12] to evaluate the performance of the regressed metrics is the PCC, we have chosen the following objective function r to be maximized within the PSO:

$$r(P) = \frac{\sum_{i=1}^N (f(LC_i) - \overline{f(LC)})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (f(LC_i) - \overline{f(LC)})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (5)$$

where the function f is the logistic transformation given by Eq. 3, $f(LC_i)$ is the logistically transformed value of the proposal LC for the i -th image of the database of N images, $\overline{f(LC)}$ and \bar{y} are the means of the respective data sets.

4 Results

Starting from Eq. 4 we have defined two different metrics LC_{BJ} and LC_{NJ} for images affected by blur artifacts, followed by JPEG compression, and images affected by noise followed by JPEG, respectively. To this end, we have considered separately each of the two parts of the MD-IVL database.

In the case of the MD-IVL-BJ we have defined the LC_{BJ} metric using $M1 - M4$ listed in Sect. 3.1. We have divided in half the dataset, obtaining two parts (MD-IVL-BJ1 and MD-IVL-BJ2), of 175 images each, corresponding to all the distorted versions of five original images. In this way the image contents present in each part of the dataset do not overlap. One image content is defined as all the distorted versions of a same original image.

We have performed 1000 runs of the PSO on MD-IVL-BJ1, to set the optimal parameters $\{a_k\}$ to obtain LC_{BJ} . The metrics were previously normalized in the range $[0, 1]$, and the search space of each parameter was set in the range $[-1, 1]$. Within the 1000 runs, the average PCC (fitness function of PSO optimization, Eq. 5) is 0.9065 with standard deviation of 0.0002. These values indicate the convergence of the sequence of solutions. The optimal parameters are thus obtained averaging the 1000 solutions. The metric LC_{BJ} reads:

$$LC_{BJ} = 1.000 \times M1 + 0.006 \times M2 + 0.829 \times M3 + 0.482 \times M4 \quad (6)$$

We observe from Eq. 6 that the highest contribution to the linear combination comes from the blurriness metric $M1$ followed by the blockiness metric $M3$. The lowest contribution comes from $M2$.

We test the performance of the proposed metric on the MD-IVL-BJ2 and on the LIVE-MD part 1 datasets. We report the corresponding PCC in Table 1 compared also to BRISQUE and NIQE metrics. In Fig. 3 we plot the logistic regression curves for the MD-IVL-BJ2 dataset, comparing our LC_{BJ} , with BRISQUE and NIQE metrics.

Table 1. Performance evaluation of the LC_{BJ} in comparison to BRISQUE and NIQE, in terms of PCC for the MD-IVL-BJ2 and LIVE-MD part 1 respectively.

Dataset	LC_{BJ}	BRISQUE	NIQE
MD-IVL-BJ2	0.8278	0.8069	0.7372
LIVE-MD part 1	0.8761	0.8687	0.9088

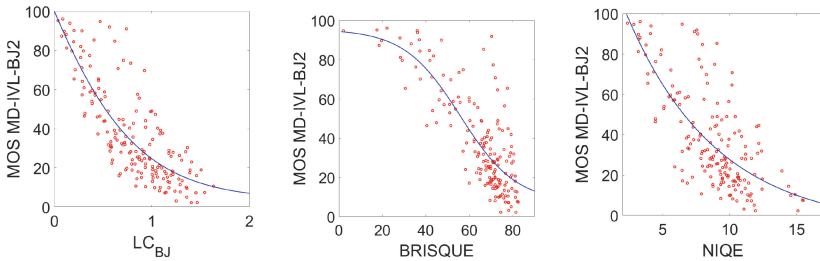


Fig. 3. Logistic regression curves for the MD-IVL-BJ dataset: LC_{BJ} , BRISQUE and NIQE.

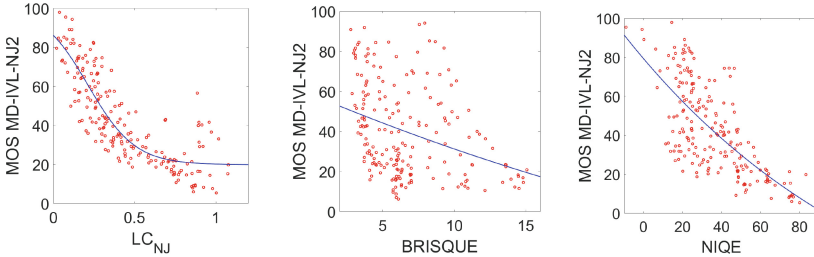
For the new dataset MD-IVL-BJ, the performance of the proposed linear combination LC_{BJ} is the highest one. Even if for the LIVE-MD data the best performance is achieved by NIQE, our proposal shows a good performance, taking into account that it is based on simple metrics signal based.

In the case of the MD-IVL-NJ we have defined the LC_{NJ} metric using metrics $M2 - M5$ listed in Sect. 3.1. As before, we have divided in half the dataset, obtaining two parts (MD-IVL-NJ1 and MD-IVL-NJ2), of 200 images each, corresponding to all the distorted versions of five original images. We have also in this case performed 1000 runs of the PSO on MD-IVL-NJ1, to set the optimal parameters $\{a_k\}$ to obtain LC_{NJ} . Within the 1000 runs, the average PCC (fitness function of PSO optimization, Eq. 5) is 0.9272 with standard deviation of 0.0006. Again, the optimal parameters are obtained averaging the 1000 solutions. The metric LC_{NJ} reads:

$$LC_{NJ} = -0.078 \times M2 + 0.3483 \times M3 + 1.000 \times M4 + 0.6717 \times M5 \quad (7)$$

Table 2. Performance evaluation of the LC_{NJ} in comparison to BRISQUE and NIQE, in terms of PCC for the MD-IVL-NJ2 dataset.

Dataset	LC_{NJ}	BRISQUE	NIQE
MD-IVL-NJ2	0.8660	0.3379	0.6960

**Fig. 4.** Logistic regression curves for the MD-IVL-NJ dataset: LC_{NJ} , BRISQUE and NIQE.

We observe from Eq. 7 that the highest contribution to the linear combination now comes from the blockiness metric $M4$ followed by the noise metric $M5$. Also in this case the lowest contribution comes from $M2$. Considering both LC_{BJ} and LC_{NJ} we can also observe that the overall contribution of the blockiness metrics is of the same order (the sum of their weighting coefficients is similar).

We test the performance of the proposed metric on the MD-IVL-NJ2, reporting the corresponding PCC in Table 2. Also the performance of BRISQUE and NIQE are included in the table for comparison. We observe that the proposal LC_{NJ} outperforms both general purpose metrics. The low value for the BRISQUE PCC is comparable with the corresponding one for the LIVE-MD database part 2 (noise + blur) [10]. Taking into account this fact, probably the lower performance showed by BRISQUE in both cases can be attributed to the presence of noise artifacts.

In Fig. 4 the corresponding logistic regression curves for the MD-IVL-NJ2 dataset, for LC_{NJ} , BRISQUE and NIQE metrics are shown.

5 Conclusions

In this work we have focused on multiply distorted image quality assessment. We have generated a database composed of two different parts. Starting from ten original color images of the SD-IVL database we have simultaneously distorted them, considering two different combinations of distortions: blur and JPEG and noise and JPEG. Psycho-physical experiments were conducted on each of these databases. In particular we have here proposed two NR-MD metrics, one for each couple of distortions, as a linear combination of state of the art NR single distortion ones. The promising results obtained suggest possible future research. Firstly, other NR single distortions metrics can be taken into account within

the linear combination. Moreover a non linear combination can be investigated using for example Genetic Programming as it was done in [24].

Acknowledgments. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

1. Ciocca, G., Corchs, S., Gasparini, F., Schettini, R.: How to assess image quality within a workflow chain: an overview. *Int. J. Digit. Libr.* **15**(1), 1–25 (2014)
2. Gabarda, S., Cristóbal, G.: Blind image quality assessment through anisotropy. *J. Opt. Soc. Am. A* **24**(12), B42–B51 (2007)
3. Choi, M., Jung, J., Jeon, J.: No reference image quality assessment using blur and noise. *Int. J. Comput. Sci. Eng.* **2**(3), 76–80 (2009)
4. Cohen, E., Yitzhaky, Y.: No-reference assessment of blur and noise impacts on image quality. *Signal Image Video Process.* **4**, 289–302 (2010)
5. Zhu, X., Milanfar, P.: Automatic parameter selection for denoising algorithms using a no-reference measure of image content. *IEEE Trans. Image Process.* **19**(12), 3116–3132 (2010)
6. Mittal, A., Moorthy, A., Bovik, A.: No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **21**(2), 4695–4708 (2012)
7. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a completely blind image quality analyzer. *IEEE Signal Process. Lett.* **20**, 209–212 (2013)
8. Chandler, D.M.: Seven challenges in image quality assessment: past, present, and future research. *ISRN Signal Processing* 23 Article ID 905685, 53 p. (2013)
9. Ciocca, G., Corchs, S., Gasparini, F., Schettini, R.: Modeling image quality, pp. 569–580. IGI Global (2014)
10. Jayaraman, D., Mittal, A., Moorthy, A.K., Bovik, A.: Objective quality assessment of multiply distorted images. In: *Proceedings of the Asilomar Conference on Signals, Systems and Computers* (2012)
11. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process.* **25**(1), 372–387 (2016)
12. VQEG: Vqeg final report of fr-tv phase ii validation test. Technical report, Video Quality Experts Group (VQEG) (2003)
13. Corchs, S., Ciocca, G., Bricolo, E., Gasparini, F.: Predicting complexity perception of real world images. *PLoS ONE* **11**(6), e0157986 (2016)
14. Corchs, S., Gasparini, F., Schettini, R.: No reference image quality classification for JPEG-distorted images. *Digital Signal Process.* **30**, 86–100 (2014)
15. ITU: Methodology for the subjective assessment of the quality for television pictures. Technical report, ITU-R Rec. BT. 500–511 (2002)
16. Marziliano, P., Dufaux, F., Winkler, S., Ebrahimi, T.: Perceptual blur and ringing metrics: application to JPEG2000. *Signal Process. Image Commun.* **19**(2), 163–172 (2004)
17. Wu, H., Yuen, M.: A generalized block-edge impairment metric for video coding. *IEEE Signal Process. Lett.* **4**, 317–320 (1997)
18. Wang, Z., Bovik, A.C., Evans, B.L.: Blind measurement of blocking artifacts in images. In: *Proceedings of International Conference on Image Processing*, vol. 3, 981–984. IEEE (2000)

19. Immerkaer, J.: Fast noise variance estimation. *Comput. Vis. Image Underst.* **64**(2), 300–302 (1996)
20. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.* **15**(11), 3440–3451 (2006)
21. Larson, E., Chandler, D.: Most apparrent distortion: full reference image quality assessmente and the role of strategy. *J. Electron. Imaging* **19**(011006), 1–21 (2010)
22. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: *Proceedings of IEEE International Conference on Neural Networks*, vol. 4, pp. 1942–1948 (1995)
23. Bianco, S., Schettini, R.: Two new von kries based chromatic adaptation transforms found by numerical optimization. *Color Res. Appl.* **35**(3), 184–192 (2010)
24. Corchs, S., Ciocca, G., Francesca, G.: A genetic programming approach to evaluate complexity of texture images. *J. Electron. Imaging* **25**(6), 061408 (2016)