Noisy images-JPEG compressed: subjective and objective image quality evaluation

Silvia Corchs, Francesca Gasparini, and Raimondo Schettini

Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy

ABSTRACT

The aim of this work is to study image quality of both single and multiply distorted images. We address the case of images corrupted by Gaussian noise or JPEG compressed as single distortion cases and images corrupted by Gaussian noise and then JPEG compressed, as multiply distortion case. Subjective studies were conducted in two parts to obtain human judgments on the single and multiply distorted images. We study how these subjective data correlate with No Reference state-of-the-art quality metrics. We also investigate proper combining of No Reference metrics to achieve better performance. Results are analyzed and compared in terms of correlation coefficients.

Keywords: image quality assessment, no reference metrics, 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.). 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.[1–3] 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.[4,5] Following Mittal et al.[4] it is also possible to individuate two subcategories of blind models: the model is called Opinion-Aware (OA) if it has been trained on a database(s) of human rated distorted images and associated subjective opinion scores, otherwise it is called Opinion-Unaware (OU). An overview of the different objective and subjective IQA methods can be found in the review article by Chandler.[6] 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. Recently Jayaraman et al.[7] has presented a database of multiply distorted images, where two scenarios are considered: images first blurred and then JPEG compressed, and images first blurred and then corrupted by white Gaussian noise.

To compare objective and subjective results different performance measures are used. The Video Quality Experts Group (VQEG)[8] 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

Send correspondence to: Silvia Corchs E-mail: silvia.corchs@disco.unimib.it

Image Quality and System Performance XI, edited by Sophie Triantaphillidou, Mohamed-Chaker Larabi,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 9016, 90160V · © 2014 SPIE-IS&T
CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2039273

transformed predicted values. Recently, it has also been proposed a Monotonic Regression.[9] This function is obtained by solving an optimization problem that yields the highest PCC and does not depend on any parameter settings.

We here address the case of images corrupted by Gaussian noise and images JPEG compressed as single distortion cases and images corrupted by Gaussian noise and then JPEG compressed, as multiple distortion case. To this end, we have generated a database (hereafter called IVL database) of images corrupted by single distortion (Gausssian noise and JPEG) and by multiple distortions (Gaussian noise followed by JPEG compression). Subjective studies were conducted on this database to obtain human judgments on both the single and multiply distorted images and the corresponding psycho-visual data were collected. We study in this work how these subjective data correlate with NR state-of-the-art metrics. Among the available metrics for single distortions, we have chosen one metric specific for noise[10] and one metric specific for JPEG-blockiness[11] that highly correlate with the corresponding subjective data. Also two general purpose metrics are taken into account: one OA[4] and the other OU.[5] We demonstrate in the experimental section that neither metrics specifically developed for single distortion nor general purpose ones are able to properly fit the subjective data in the case of multiple distortion noise-JPEG. Up to our knowledge, there exist no NR methods to assess image quality for simultaneous noise and JPEG artifacts. To this end, we here propose linear combinations of the considered NR metrics where the weighting coefficients are obtained using a particle swarm optimization.[12,13] In the experimental section, the performance of the NR metrics considered are compared with the performance obtained with the proposed optimized linear combinations.

2. SUBJECTIVE DATA: THE IVL DATABASE

The IVL database originates from 20 reference images of 886x591 pixels (15x10 cm at 150 dpi, typical printing parameters for natural photos), 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 Figure 1.

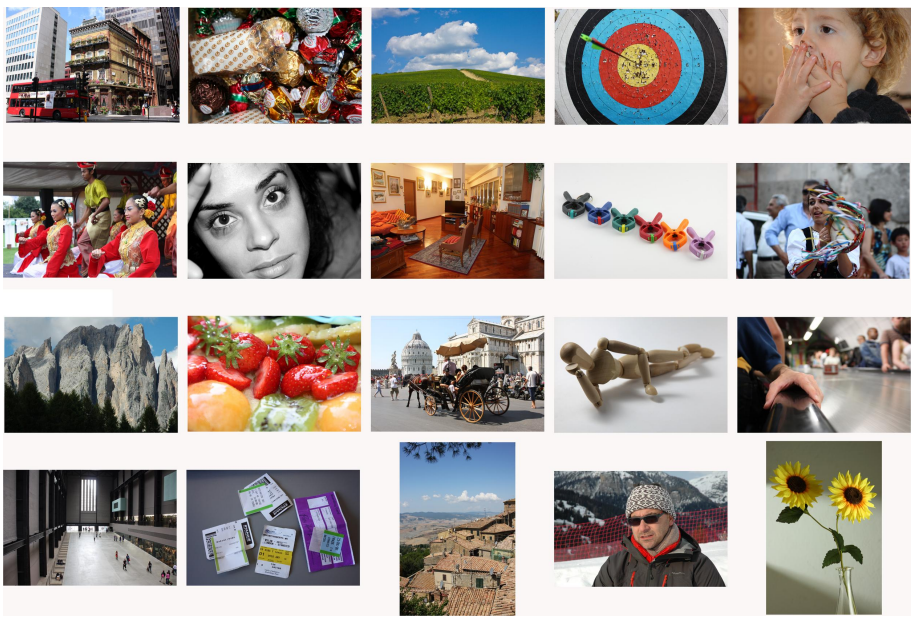


Figure 1. The 20 reference images of the IVL database.

Starting from these images we have generated:

- A database of 200 noisy images. The distorted images have been obtained as follows: for each of the 20 reference images we have created 10 corrupted versions with: 1, 2, 3, 4, 5, 6, 8, 10, 12 and 14 gray levels of standard deviation on the luminance channel.

- A database of 180 JPEG compressed images. The compressed images were generated using the Matlab imwrite function. As the Q-factor depends on the specific JPEG compression algorithm used, we have adopted the bit per pixel (bpp) Ratio (*bppR*) with respect to a reference, finding iteratively the Q-factors that better match the corresponding *bppR* values. As reference we have adopted the $Q = 100$ compressed image, where the compression is mainly due to the sub sampling of the chroma channels and to lossless algorithms. For each of the 20 original images, we have created 9 compressed versions with the following *bppR*: $1(Q = 100), 0.707, 0.5, 0.25, 0.177, 0.125, 0.105, 0.088, 0.0625$.
- A database of 800 multiply distorted images. These distorted images have been generated as follows: each of the 200 noisy images were further processed by 4 different levels of JPEG compression, corresponding to Q factor values of 100, 50, 30, and 10.

we show in Figure 2 a reference image (a) together with the most distorted versions corresponding to the single noise (b) and JPEG (c) distortions and the multiply distorted noise-JPEG one (d).

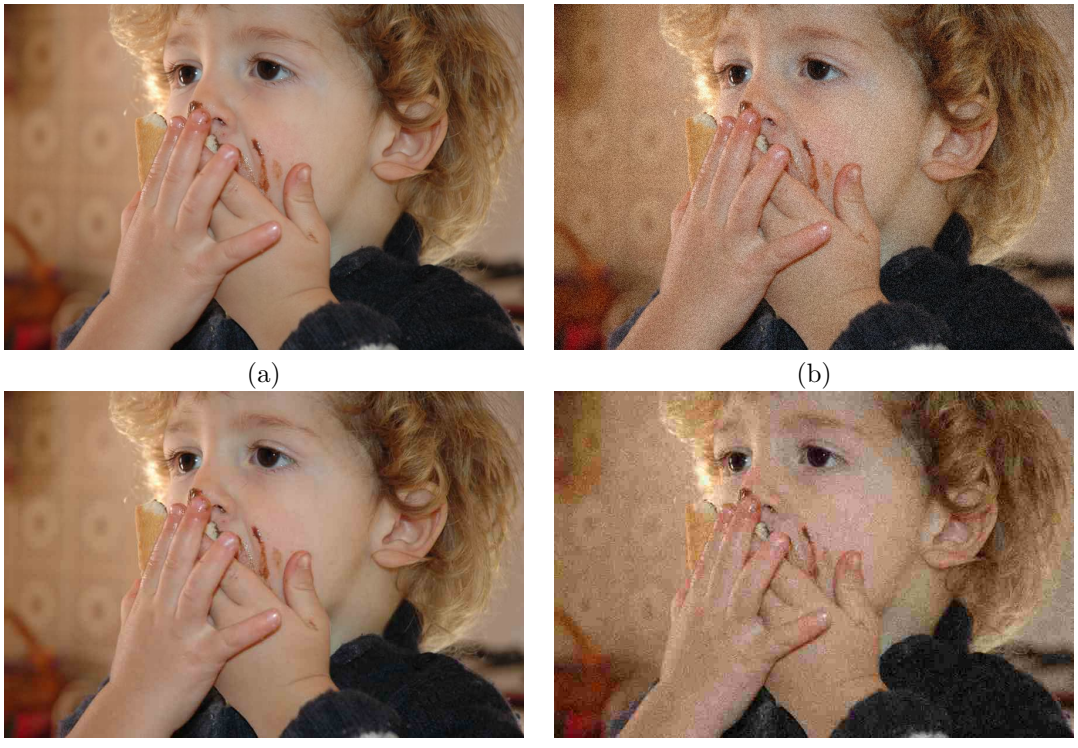


Figure 2. a) A reference image from IVL database and and its most distorted versions for single distortion b) noise and c) JPEG) and d) multiple JPEG-noise artifacts.

For collecting the subjective data on these three different databases, we have adopted a Single Stimulus method (SS),[14] where all the images are individually shown. We have decided to adopt the SS method to better represent the reality where users of digital photographs do not in general dispose of reference images (NR IQA). 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.[14]

For the two single distortion databases (noise and JPEG) all the distorted images (200 and 180 respectively) have been assessed, while for the multiply distorted database we present here results corresponding to a subset of 400 multiply distorted images.

3. OBJECTIVE DATA

The subjective scores described in Section 2, collected in terms of Mean Opinion Scores (MOS) have been correlated with different NR metrics, listed in the next subsections.

The metrics and the subjective scores (of a given database) have been correlated using a logistic function. Denoting by y_i the MOS value of the $i-th$ image of the database ($i = 1, ...N$ with N the total number of distorted images) and by x_i the corresponding objective metric value, the logistic transformation reads:

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

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

- The noise specific NR metric by Immerkaer[10] (hereafter called IMMERKAER) estimates the standard deviation of Additive White Gaussian Noise (AWGN) from a single image using a Laplacian mask filtering approach. The metric implementation by Foi[15] is used in the present work.
- The JPEG-blockiness metric by Wang et al.[11] (hereafter called WBE) is formulated in the frequency domain and models the blocky image as a non-blocky image interfered with a pure blocky signal. The goal of the blocking effect measurement algorithm is then to detect and estimate the power of the blocky signal. Luminance and texture masking effects are also integrated within the metric.

3.2 NR General Purpose metrics

- The general purpose metric BRISQUE by Mittal et al.[4] employs statistics measured in the spatial domain. BRISQUE operates on two image scales; for each scale, different statistical features are extracted to be used within a two stage classification and regression framework. It uses scene statistics of locally normalized luminance coefficients to quantify possible losses of naturalness in the image due to the presence of distortions, leading to a holistic measure of quality. The authors provide the algorithm implementation where LIVE database[16] has been used for the training.
- The general purpose metric NIQE by Mittal et al.[5] is based on constructing a collection of quality aware features and fitting them to a Multivariate Gaussian (MVG) model. The quality aware features are derived from a simple but highly regular Natural Scene Statistic (NSS) model. The quality of a given test image is then expressed as the distance between the MVG fit of the NSS features extracted from the test image, and a MVG model of the quality aware features extracted from the corpus of natural images.

3.3 NR metric for noisy images-JPEG compressed: our proposal

Up to our knowledge, there exists no specific distortion NR metric that takes into account simultaneously noise and JPEG artifacts. To this end, in our paper we propose to adopt a linear combination of the metrics presented in Section 3.1.

This combination can be written as follows:

$$M = a \times WBE + b \times IMMERKAER + c \times BRISQUE + d \times NIQE \tag{2}$$

With respect to this general equation we consider four possible versions, combining only subsets of these metrics, corresponding to:

- $M1 = a \times WBE + b \times IMMERKAER + c \times BRISQUE + d \times NIQE$, i.e. all the metrics are considered;
- $M2 = a \times WBE + b \times IMMERKAER + d \times NIQE$, i.e. the two distortion specific metrics are considered together with the NIQE general purpose one;

- $M3 = a \times WBE + b \times IMMERKAER + c \times BRISQUE$, i.e. the two distortion specific metrics are considered together with the BRISQUE general purpose one;
- $M4 = a \times WBE + b \times IMMERKAER$, i.e. only the two distortion specific metrics are considered.

For each of the proposals, the set of optimal parameters P_{opt} are found using Particle Swarm Optimization (PSO)[12,13] over the set $P \in \Re^n$ of feasible solutions.

Recalling that one of the criteria recommended by the VQEG[8] to evaluate the performance of the regressed metrics is the PCC, we have chosen the following objective function r to be maximized:

$$r(P) = \frac{\sum_{i=1}^{N}(f(M_i) - \overline{f(M)})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(f(M_i) - \overline{f(M)})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{3}$$

where the function f is the logistic transformation given by Equation 1, $f(M_i)$ is the logistically transformed value of the proposal M for the $i-th$ image of the database of N images, $\overline{f(M)}$ and $\bar{y}$ are the means of the respective data sets. In Equation 3, M indicates any of our four proposals ($M1$, $M2$, $M3$, and $M4$).

The P_{opt} parameter values are obtained as:

$$P_{opt} = \max_{P \in \Re^n}(r(P)). \tag{4}$$

4. RESULTS

In this section we evaluate the performance the correlation between subjective and objective data of the IVL database, in the following cases:

- *Single Distortion Noise*: considering the three NR metrics IMMERKAER, BRISQUE, and NIQE;
- *Single Distortion JPEG*: considering the three NR metrics WBE, BRISQUE, NIQE;
- *Multidistortion MD*: considering the four NR metrics WBE, IMMERKAER, BRISQUE, and NIQE, and our 4 proposals M1, M2, M3, and M4.

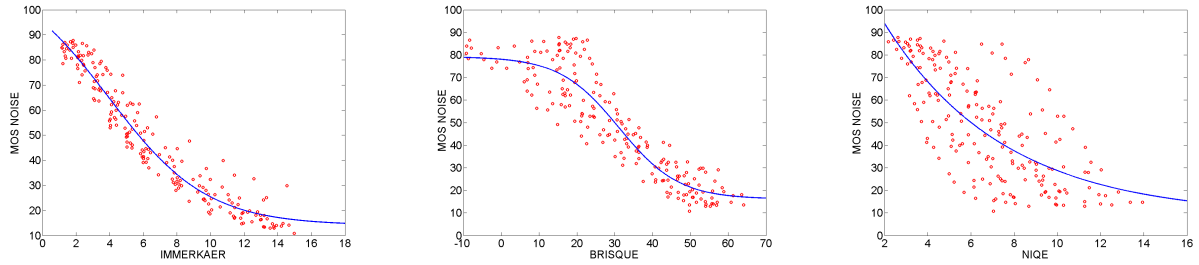
These performances are evaluated in terms of the PCC, SROCC and RMSE.

In Figure 3 we plot the MOS versus IMMERKAER, BRISQUE and NIQE metrics and the corresponding logistic regression curves for the noisy distorted images. The statistic correlation coefficients PCC, SROCC and RMSE are reported in Table 1. Recalling that values of 1 for both PCC and SROCC mean a perfect linear correlation, from Table 1 it comes out that the metric by Immerkaer is the one that best correlates with the subjective scores.

NOISE	IMMERKAER	BRISQUE	NIQE
SROCC	0.9660	0.9096	0.7300
PCC	0.9688	0.9262	0.7393
RMSE	5.850	8.9033	15.909

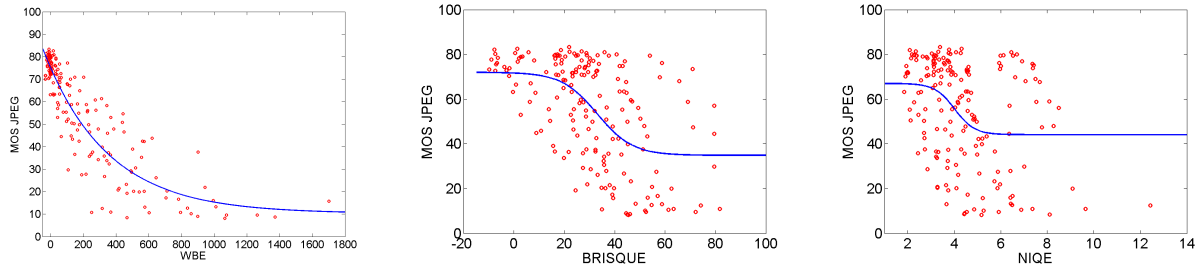
Table 1. Performance evaluation of the IMMERKAER, BRISQUE and NIQE metrics on noisy images, in terms of correlation coefficients and RMSE.

In Figure 4 the logistic regressions obtained for the WBE, BRISQUE and NIQE metrics on the JPEG distorted images are shown. The corresponding correlation coefficients and RMSE are reported in Table 2. The WBE



(a) (b) (c)

Figure 3. Logistic regression curves for the noise distorted images of the IVL database. a)IMMERKAER, b) BRISQUE, and c) NIQE.



(a) (b) (c)

Figure 4. Logistic regression curves for the JPEG distorted images of the IVL database. a)WBE, b) BRISQUE, and c) NIQE.

metric is the one that best correlates with the subjective data. The lower performance of BRISQUE could be partially attributed to the fact that it is an OA method and it has been trained on the LIVE data.

In Figure 5 the subjective scores for the MD images are plotted versus WBE, IMMERKAER, BRISQUE and NIQE metrics. Note that for this database these plots are more spread than in the previous cases. The general purpose metrics applied to MD data seem to be more suitable than single distortion ones. In fact, for WBE and IMMERKAER (that evaluate single distortions), a severe non-monotone behavior of the metric response is observed near the y-axis (see Figures 5 a and b). For example in case of WBE, the metric gives values near zero for images not compressed, while the MOS are spread along the y-axis due to the presence of different levels of noise. As expected the correlation performances of the WBE is much lower than in the single distortion experiment (compare Table 3 with Table 2). Instead, for IMMERKAER it is not possible to find a feasible logistic regression in the MD case.

In Table 4 the parameters of Equation 2 found after optimization for the different combinations proposed $M1 - M4$ are reported. Since the WBE metric predictions can assume absolute values in the order of 1000, the parameter a is indicated as a ratio of 100. In Figure 6 the MOS MD versus each of the combinations proposed are shown. Finally, in Table 5 the proposal performances are evaluated. Comparing PCC, SROCC and RMSE of Tables 5 and 3 we observe that all the combinations proposed outperform the performances of the metrics WBE, BRISQUE and NIQE.

JPEG	WBE	BRISQUE	NIQE
SROCC	0.8922	0.5386	0.3536
PCC	0.9059	0.5578	0.3830
RMSE	9.7721	19.1473	21.3109

Table 2. Performance evaluation of the WBE, BRISQUE and NIQE metrics on JPEG distorted images, in terms of correlation coefficients and RMSE.

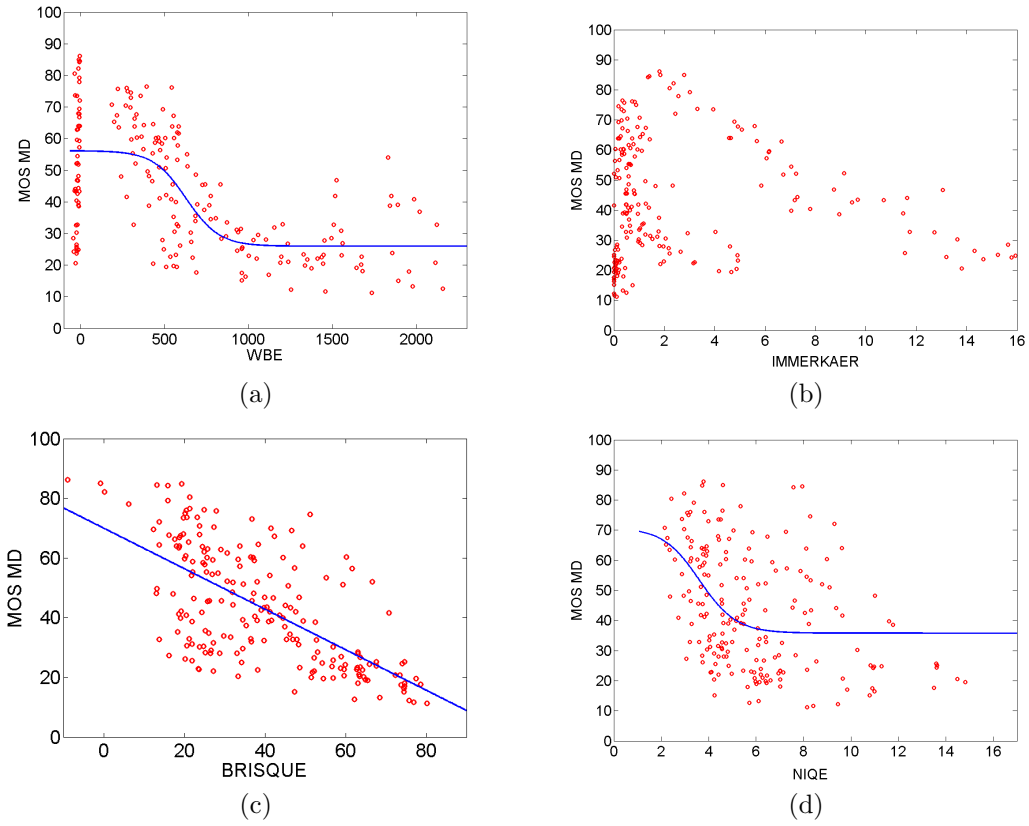


Figure 5. Logistic regression curves for the MD images of the IVL database. a)WBE, b)IMMERKAER, c) BRISQUE, and d) NIQE.

MD	WBE	IMMERKAER	BRISQUE	NIQE
SROCC	0.5575	-	0.6541	0.4371
PCC	0.6515	-	0.6583	0.4407
RMSE	15.8242	-	14.7086	17.5400

Table 3. Performance evaluation of the WBE, IMMERKAER, BRISQUE and NIQE metrics on MD images in terms of correlation coefficients and RMSE.

5. CONCLUSIONS

In this work we have focused on multiply distorted image quality assessment. We have generated a database of distorted images for single noise, single JPEG and simultaneous noise and JPEG artifacts. Psychovisual experiments were conducted on each of these databases. The subjective scores have been correlated with different NR metrics. In particular we have shown that distortion-specific metrics (designed for measuring noise or JPEG

Weights	WBE a	IMMERKAER b	BRISQUE c	NIQE d
M1	0.75/100	0.59	0.014	0.14
M2	1.00/100	0.97	0	0.21
M3	1.00/100	0.81	0.024	0
M4	1.00/100	0.77	0	0

Table 4. The parameters of Equation 2 found after optimization for the different combining proposals $M1 - M4$.

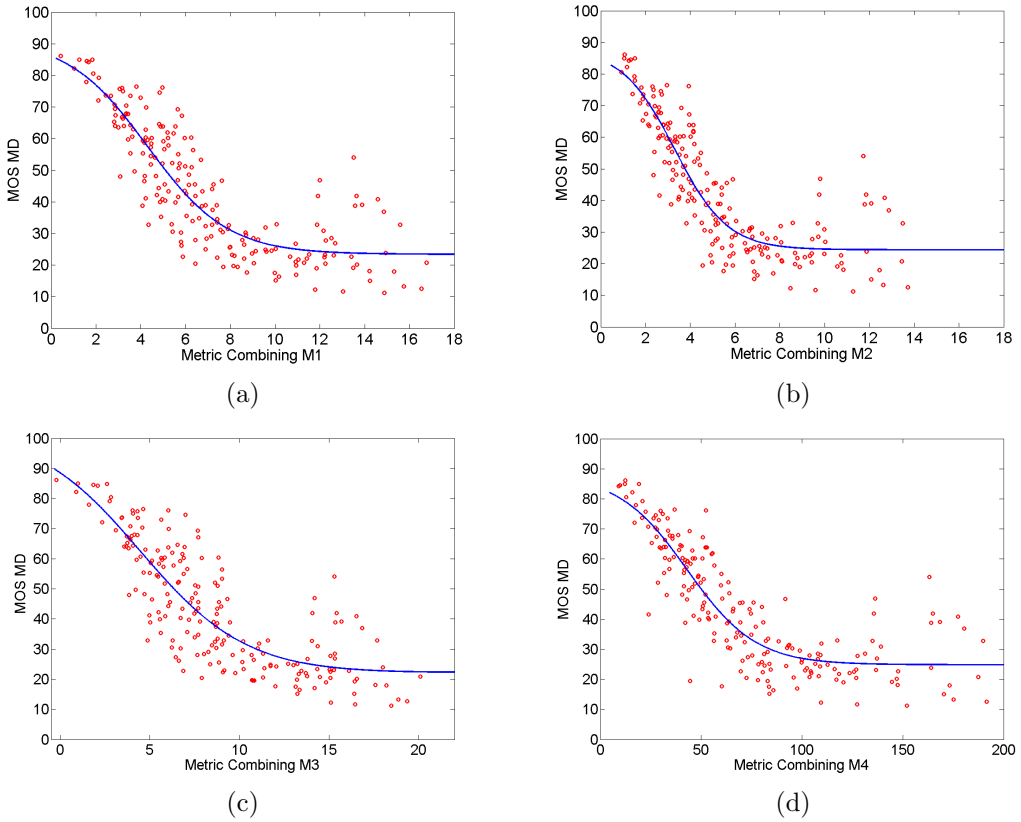


Figure 6. Logistic regression for MD database and each of the proposals $M1 - M4$.

MD	M1	M2	M3	M4
SROCC	0.8525	0.8396	0.8460	0.8470
PCC	0.8872	0.8838	0.8833	0.8839
RMSE	9.0130	9.1418	9.1617	9.1364

Table 5. Performance evaluation of our four proposals in case of MD images, in terms of correlation coefficients and RMSE.

artifacts) are not able to predict the subjective scores in case of noisy images-JPEG compressed. General purpose metrics seem to be more suitable, even if the correlation performances are still very low. We have also proposed different linear combination of these NR distortion-specific and general purpose metrics. The optimized weights were derived from the particle swarm optimization method. In general, the different combinations proposed show good performance when correlated with the subjective scores of the multiply distorted images in terms of correlation coefficients (PCC and SROCC) and RMSE and outperform both distortion specific and general purpose metrics.

REFERENCES

[1] Gabarda, S. and Cristóbal, G., "Blind image quality assessment through anisotropy," *J. Opt. Soc. Am. A* **24**, B42–B51 (Dec 2007).

[2] Choi, M., Jung, J., and Jeon, J., "No reference image quality assessment using blur and noise," *International Journal of Computer Science and Engineering* **2**(3), 76–80 (2009).

[3] Cohen, E. and Yitzhaky, Y., "No-reference assessment of blur and noise impacts on image quality," *Signal, Image and Video Processing* **4**, 289–302 (2010).

[4] Mittal, A., Moorthy, A., and Bovik, A., "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing* **21**(2), 4695–4708 (2012).
[5] Mittal, A., Soundararajan, R., and Bovik, A. C., "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters* **20**, 209–212 (2013).
[6] Chandler, D. M., "Seven challenges in image quality assessment: Past, present, and future research," *ISRN Signal Processing* **23**, Article ID 905685, 53 pages (2013).
[7] D. Jayaraman, A. Mittal, A. M. and Bovik, A., "Objective quality assessment of multiply distorted images," in [*Proc. of the Asilomar Conference on Signals, Systems and Computers*], (2012).
[8] VQEG, "Vqeg final report of fr-tv phase ii validation test," tech. rep., Video Quality Experts Group (VQEG) (2003).
[9] Han, Y., Cai, Y., Cao, Y., and Xu, X., "Monotonic regression: A new way for correlating subjective and objective ratings in image quality research," *IEEE Transactions on Image Processing* **21**, 2309–2313 (2012).
[10] Immerkaer, J., "Fast noise variance estimation," *Computer Vision and Image Understanding* **64**(2), 300 – 302 (1996).
[11] Wang, Z., Bovik, A. C., and Evans, B. L., "Blind measurement of blocking artifacts in images," in [*Proc. International Conference on Image Processing*], **3**, 981–984, IEEE (2000).
[12] Bianco, S. and Schettini, R., "Two new von kries based chromatic adaptation transforms found by numerical optimization," *Color Research & Application* **35**(3), 184–192 (2010).
[13] J., K. and R., E., "Particle swarm optimization," in [*Proc IEEE Int Conf Neural Networks*], **4**, 1942–1948 (1995).
[14] ITU, "Methodology for the subjective assessment of the quality for television pictures," tech. rep., ITU-R Rec. BT. 500-11 (2002).
[15] Foi, A., "Anisotropic nonparametric image restoration demobox," http://www.cs.tut.fi/ lasip/2D/ (2006).
[16] Sheikh, H., Z.Wang, Cormack, L., and Bovik, A., "Live image quality assessment database release 2," http://live.ece.utexas.edu/research/quality (2005).