International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 13 (2021) pp.348-363
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Received: 21 August 2021; Accepted: 11 Nov. 2021; Publish: 9 Dec. 2021

Strategy for Selecting a Quality Index for Images

María Lucía Pappaterra[1], Silvia María Ojeda[2], Marcos Alejandro Landi[3] and Ronny Obed Vallejos[4]

[1] Universidad Nacional de Córdoba, Facultad de Matemática, Astronomía, Física y Computación,
Probability and Statistics Department, CIEM-CONICET
Av. Medina Allende 2144, zip code 5000, Córdoba, Argentina
lucia.pappaterra@unc.edu.ar

[2] Universidad Nacional de Córdoba, Facultad de Matemática, Astronomía, Física y Computación,
Probability and Statistics Department, CIEM-CONICET
Av. Medina Allende 2144, zip code 5000, Córdoba, Argentina
sm.ojeda@unc.edu.ar

[3] Universidad Nacional de Córdoba, Facultad de Ciencias Exactas, Físicas y Naturales,
Av. Vélez Sarsfield 1611, zip code 5000, Córdoba, Argentina
marcoslandi1980@gmail.com

[4] Universidad Técnica Federico Santa María, Department of Mathematics,
Av. España 1680, Valparaíso, Chile
ronny.vallejos@usm.cl

Abstract: **In the past few decades, many image quality indices have been developed. However, they stem from different theoretical frameworks, application scenarios and purposes. Thus, users and researchers are often faced with the time-consuming task of deciding which quality index to choose when they require a reliable image quality index that is capable of emulating the human visual system (HVS). In this work, general criteria for selecting the most appropriate index from a given set of quality indices according to application needs are established. These criteria are based on the statistical coefficients of correlation and concordance. It is discussed why Kendall's Tau and Spearman's rank correlation coefficients—which are widely used to compare quality index performance—are not sufficient for this purpose; moreover, additional nonparametric tests and methods of agreement are incorporated: the concordance coefficients (Kendall's w, Cohen's kappa, Scott's pi and Fleiss' kappa) not explored so far, to determine the best procedures to compare digital images. The combination of all these strategies led to a more complete comparison method, from which a ranking of quality indices could be generated from any set of them. As an application, the performance and suitability of a large number of quality indices for various real-world scenarios is compared. Our experiments reveal that the indices are sensitive to the type of distortions. This work expanded previous studies by incorporating directional indices, which perform well in the numerical experiments developed using real datasets.**

Keywords: Image analysis, Image quality indices, Measures of association, Concordance coefficients, Distortion types, Machine vision and scene understanding.

I. Introduction

The use of digital images as a convenient mechanism for representing information has rapidly increased in the last decades [1]. Consequently, a large number of researchers and practitioners have focused their efforts on methodological aspects and the development of algorithms for the analysis and processing of images with applications in a variety of different fields. Because image processing is commonly subject to errors, due to acquisition, discretization, compression and transmission [2], these efforts are mainly focused in improving the appearance of images of interest. In this framework, automatic and effective mechanisms to detect and measure the levels of distortions in images, quantifying their quality, are essential. There is a need to compare the performance of different image processing algorithms by quantifying and comparing the quality of their output images. Image quality assessment (IQA) has been a very useful approach to this objective. This topic can be classified as subjective or objective IQA [1].

Subjective IQA—based on the average opinion of a set of observers—provides the most accurate measures of quality because the human eye is the final receptor of all visual communication systems [3]. One of the most significant contributions of subjective IQA in recent decades has been the construction of databases consisting of digital images featuring various types of distortions labeled with a mean opinion score (MOS)—a subjective rating obtained based on experiments involving human observers. The abovementioned databases include the Tampere Image Database 2008 (TID2008) [4], the Tampere Image Database 2013 (TID2013) [5] [6], and the LIVE Database available at `http://live.ece.utexas.`

edu/research/quality, among others.

Although the most accurate approach, subjective IQA is often slow, expensive and laborious, making it unsuitable for quantifying image quality or for assessing image similarity within visual communication systems. [1]. This shortcoming has motivated the exploration of objective IQA indices, seeking alternative ways to assess image quality. The goal of this approach is to design tools that can accurately and automatically quantify image quality. Objective IQA is divided into three main research areas based on the availability of reference images: full-reference IQA (FR-IQA) ([7], [4], [8]), reduced-reference IQA (RR-IQA) ([9], [10]), and no-reference IQA (NR-IQA) [11], [12]). In the FR-IQA context, a quality index M is a function that quantifies the quality of an image I with respect to a reference image I_R and is denoted as $M(I, I_R)$ [13]. These quality indices rely on high-quality images being available as references to assess the differences between them and their distorted versions, hence verifying the importance of image databases such as those previously mentioned.

Researchers are often faced with the dilemma of deciding which FR-IQA index to choose from. Some studies have been carried out in this regard [6] [14] [15] [16] and have been precursors in the literature. The aim of this paper is to provide a methodology for selecting an appropriate mechanism to quantify image quality according to application needs; the objective is to answer the following question: which mechanism should be chosen, and why?

In this work statistical methods were used to carry out a comparative analysis between a subset of FR-IQA indices while suggesting criteria of which one among them should be selected according to its performance in different applied scenarios. One important work in this respect is the study carried out by Ponomarenko et al. [6]. They developed a comparative study between objective IQA indices using the nonparametric correlation coefficients of Spearman and Kendall. In this paper, the statistical tools used in [6] to compare FR-IQA indices are extended, exploring new comparison methods. In addition to the Spearman and Kendall nonparametric correlation coefficients, this analysis includes Kendall's w multivariate correlation—a generalization of Spearman's coefficient—as well as Cohen's kappa, Scott's pi and Fleiss' kappa concordance coefficients—all calculated based on confusion matrices [17] [18] [19]. The advantage of incorporating these procedures is that they provide a more comprehensive understanding of the relationship between each index and the MOS and among a set of indices and the MOS. The inclusion of multivariate coefficients makes the comparisons more flexible and complete, which is a significant improvement in this comparison strategy. In addition, the concept of concordance is included. This concept, despite being similar to correlation, is a notion of agreement between two items. It is discussed why a comparison between the MOS and a given quality index based solely on correlation coefficients is incomplete, and why a more suitable statistical procedure to analyze and compare a set of FR-IQA indices should be considered. The combination of all these strategies led to a more complete comparison method, from which a ranking of quality indices can be generated.

Three directional quality indices are incorporated into the study—the *CQ index*, *gradient magnitude similarity mean (GMSM)* and *gradient magnitude similarity deviation (GMSD)*—due to their good performance reported in the literature [20, 21]. It must be emphasized that the main aim of the paper is to propose a more systematic procedure for comparing and evaluating FR-IQA indices, which can be replicated for an arbitrary set of indices.

The paper is organized as follows. In Section 2, related work is discussed and the objective IQA indices considered in our analysis are introduced. In addition, the TID2013 image database, which will be used for our numerical experiments, is also briefly described. In Section 3 statistical comparison methods are discussed in more detail and a new method is proposed. Section 4 presents the results of the numerical experiments. Finally, Section 5 includes the main conclusions of the study and a brief outline of directions for future research.

II. Related Works

A. *Quality indices comparison*

The contributions to FR-IQA are numerous ([22], [23], [3], [24], [25], [26], [27], [28], [29], [30]), however, few studies compare the performance of existing indices. Among these, we can highlight the studies carried out in [6] and [31].

In [6] Ponomarenko et al. presented the TID2013 base, intended for evaluation of FR-IQA metrics. The availability of MOS allows the use of the designed database as a fundamental tool for assessing the effectiveness of quality indices. An analysis of the correlation between MOS and a wide set of existing metrics is carried out and a methodology for determining drawbacks of existing quality indices is described. This comparative study is based on the nonparametric correlation coefficients of Spearman and Kendall. These correlation indices have been obtained both considering the full set of distorted images and specific image subsets. In this way, the performance of the different quality indices is considered according with the type of distortion that has affected the image.

Moreover, Ding et al. [31] performed a large-scale comparison of a set of FR-IQA indices in terms of their use as objectives for the optimization of image processing algorithms. Specifically, they used eleven full-reference IQA models to train deep neural networks for four low-level vision tasks: denoising, deblurring, super-resolution, and compression. They tested the models on recovering a reference image from a given initialization by optimizing the model-reported distance to the reference. They reported that for many IQA methods, the optimization does not converge to the reference image and can generate severe distortions. Subjective testing on the optimized images allowed them to rank the competing models in terms of their perceptual performance, elucidate their relative advantages and disadvantages in these tasks, and propose a set of desirable properties for incorporation into future IQA models.

Other works that can be mentioned in this regard are [15], [16], [32], [33], [34], [35].

B. Full-reference image quality indices

In this section, a brief description of the quality indices included in this study is presented. Defining these indices will help readers to better understand the research context. Computational routines are available in the repository `https://github.com/lucia15/IQA`

1) Mean squared error (MSE), signal-to-noise ratio (SNR) and related indices

One of the simplest and widely used indices for computing the quality of an image is the *mean squared error (MSE)* [14], which measures the average intensity difference between a distorted image (image y) and a reference image (image x):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2, \tag{1}$$

where x_i and y_i indicate the intensity of images x and y in the position or pixel i, respectively, and N is the total number of pixels in both images. This index does not use the structural and how they correlate.

One index related to the MSE is the *signal-to-noise ratio (SNR)*, defined as the ratio between the signal and noise power.If a distorted image y is obtained from a reference image x, by adding noise r to x, then $y = x + r$. Consequently, SNR is defined through

$$\text{SNR} = 10\log_{10}\left(\frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} r_i^2}\right). \tag{2}$$

Because the signal power is equivalent to the norm of the reference image, The SNR in (2) is equivalent to

$$\text{SNR} = 10\log_{10}\left(\frac{|x|^2}{N \cdot \text{MSE}}\right), \tag{3}$$

where

$$|x|^2 = \sum_{i=1}^{N} x_i^2.$$

$|x|^2$ can be seen as the mean square deviation of the signal x with respect to zero.

The *peak signal-to-noise ratio (PSNR)* is defined in a similar way as the SNR but using the square of the maximum intensity of signal L, instead of the mean of the square of each pixel, the PNSR is

$$\text{PSNR} = 10\log_{10}\left(\frac{L^2}{\text{MSE}}\right). \tag{4}$$

Finally, the *weighted signal-to-noise ratio (WSNR)* is defined as

$$\text{WSNR} = 10\log_{10}\left(\frac{\sum_{i=1}^{N} w_i x_i^2}{\sum_{i=1}^{N} w_i (x_i - y_i)^2}\right), \tag{5}$$

where w_i are weights according to an HVS frequency response model based on a low-contrast sensitivity function (CSF) [36].

Indices based on the MSE can be easily interpreted and are simple to compute but have several limitations that might explain why its correlation with the MOS is poor. One such limitation is these indices ignore the location of the pixels in the image and the effect that this spatial distribution can have on the perception of image quality [14]. Furthermore, the measurements derived from the SNR assume that the distortion is only caused by additive noise, independent of the signal, which makes its application problematic when there are other sources of distortion. Finally, the HVS model used in the WSNR approach is linear and spatially invariant; therefore, it cannot quantify the nonlinear and spatially variable effects of HSV, such as the well-known contrast masking effect [37].

2) Noise quality measure (NQM)

The *noise quality measure (NQM)* quantifies the impact of frequency distortion and noise injection in image restoration on the human visual system (HVS) when considering the degradation of these two sources separately.

[37] presented this index in which a distorted image is modeled using a linear frequency distortion and an additive noise injection. To yield the simulated images, nonlinear space-frequency processing is performed based on Peli's contrast pyramid [37]. The SNR is then computed for the difference between the two simulated images as a measure of image quality. Subsequently, the NQM index is

$$\text{NQM} = 10\log_{10}\left(\frac{\sum_i \sum_j O_s^2(i,j)}{\sum_i \sum_j (O_s(i,j) - I_s(i,j))^2}\right), \tag{6}$$

where $O_s(x, y)$ denotes the simulated version of the model restored image, and similarly, $I_s(x, y)$ denotes the restored image.

The authors formulate a nonlinear quasi-local processing model of the HVS by modifying Peli's contrast pyramid to measure the variation in contrast sensitivity with distance, image dimensions, and spatial frequency; the variations in the local luminance mean; the contrast interaction between spatial frequencies; and contrast masking effects.

Unlike the NQM, indices based on the SNR and linear HVS models do not account for frequency distortion and ignore the essential nonlinear processing of the HVS in the spatial and frequency domains. Furthermore, the authors demonstrated through several experiments, that the NQM performs better than the PSNR and other measurements based on linear HVS models when the images are distorted by additive noise [37].

3) Structural similarity (SSIM) index and related indices

The *SSIM* is based on the hypothesis that human visual perception is strongly adapted to extract structural information from a scene; therefore, a measure of structural similarity is a reasonable good approximation of perceived quality in an image [38]. The SSIM is defined as the product of luminance, contrast and structural (correlation) comparison between a reference image x and a distorted image y:

$$\text{SSIM}(x, y) = l(x, y) \cdot c(x, y) \cdot s(x, y), \tag{7}$$

where

$$l(x,y)=\left(\frac{2\bar{x}\bar{y}+c_1}{\bar{x}^2+\bar{y}^2+c_1}\right),$$
$$c(x,y)=\left(\frac{2s_xs_y+c_2}{{s_x}^2+{s_y}^2+c_2}\right),$$
$$s(x,y)=\left(\frac{s_{xy}+c_3}{s_xs_y+c_3}\right),$$

with $\bar{x}$, $\bar{y}$, s_x^2, s_y^2 and s_{xy} representing the sample means of x and y, the sample variances of x and y, and the sample covariance between x and y, respectively. The parameters α, β and γ are fixed and associated with the weight of each coefficient has in the final product; here, for simplicity consider $\alpha=\beta=\gamma=1$. The constants c_1, c_2, and c_3 are all nonnegative and can be settled down in such a way that preserves the definition of the SSIM index when the denominators are close to zero. Commonly, these constants are small real numbers to avoid instability when $\bar{x}+\bar{y}$ is close to zero [39]. When $c_1=c_2=0$, this index is known as the *universal quality index (UQI)* first studied in [38].

Another form of the SSIM, called the *multiscale SSIM (MSSIM)*, is conducted over multiple scales through a process of multiple stages of subsampling [40]. The SSIM index is a special case of the MSSIM using a single scale. Since the introduction of the SSIM [40], many extensions have been published and discussed [14]. Some of them include visual information fidelity [15], the visual signal-to-noise ratio [41], the most apparent distortion measure [7], the information content-weighted method [42], the feature similarity index [43], the SSIM-motivated rate-distortion optimization for video coding [44], and the perceptual quality assessment for multi-exposure image fusion [45].

4) Similarity index based on the codispersion coefficient

Ojeda et al. [20] proposed using the spatial codispersion coefficient instead of the correlation coefficient in the definition of the SSIM index. This initiative resulted in a promising new objective IQA index, called the *CQ index*, which is able to capture spatial correlation in a particular direction in a two-dimensional space.

The cross-variogram $\gamma(h)$ of the weak stationary processes $X(t)$ and $Y(t)$, with $t\in D\subset\mathbb{Z}^d$, $d\in\mathbb{N}$, is defined as

$$\gamma(h)=\mathbb{E}[(X(t+h)-X(t))(Y(t+h)-Y(t))], \quad (8)$$

where $t,t+h\in D$. The codispersion coefficient is a normalization of $\gamma(h)$, defined as [46]

$$\rho(h)=\frac{\gamma(h)}{\sqrt{V_X(h)V_Y(h)}}, \quad (9)$$

where $V_X(h)=\mathbb{E}(X(t+h)-X(t))^2$ and equivalently for $V_Y(h)$. Note that both $\gamma(h)$ and $\rho(h)$ depend on $X(t)$ and $Y(t)$, although for simplicity, only h appears in the notation in the same way the variogram of a process is defined in the literature [47]. Similar to the correlation coefficient, the codispersion coefficient satisfies that $|\rho(h)|\leq 1$.

The sample codispersion coefficient is given by

$$\hat{\rho}(h)=\frac{\sum_{t,t+h\in D'}a_tb_t}{\sqrt{\hat{V}_X(h)\hat{V}_Y(h)}}, \quad (10)$$

with $t=(t_1,t_2)$, $h=(h_1,h_2)$, $D'\subseteq D$, $\#D'<\infty$, $a_t=X(t_1+h_1,t_2+h_2)-X(t_1,t_2)$, $b_t=Y(t_1+h_1,t_2+h_2)-Y(t_1,t_2)$, $\hat{V}_X(h)=\sum_{t,t+h\in D'}a_t^2$ and $\hat{V}_Y(h)=\sum_{t,t+h\in D'}b_t^2$.

In [48], the authors demonstrated that under certain regularity conditions, $\hat{\rho}(h)$ is consistent and asymptotically normal, statistical properties that allow the construction of confidence intervals and hypothesis tests for the codispersion coefficient. All the previous indices defined in this paper are able to capture only the linear association between two given sequences. However, the codispersion coefficient captures the spatial association between images $x=\{X(t):t=1,2,...,N\}$ and $y=\{Y(t):t=1,2,...,N\}$ with respect to separation vector h.

The CQ index is a generalization of the SSIM index defined in (7) and accounts for the spatial association in a specific direction h when replacing $\hat{\rho}(h)$ in the structural part, yielding

$$\mathrm{CQ}_h(x,y)=l(x,y)\cdot c(x,y)\cdot\hat{\rho}(h), \quad (11)$$

where l and c are as in (7). This index allows for the detection of hidden similarity between a degraded image and the original image in direction h. In addition, (11) has interesting mathematical properties investigated in [20, 49]. Another extension of this type of coefficient is the CQ-max index, which is obtained by evaluating the CQ coefficient in a certain set of directions of interest [22].

5) Gradient magnitude similarity indices

Image gradient plays a very important role in the understanding of visual signals, which is used to carry structural scene information. As such, it is a crucial feature in the development of objective quality assessment indices that largely base their measurement on the preservation of this information from the original image into the test image.

Given an image, the magnitude of the gradient (G) is defined as

$$G=\sqrt{G_h^2+G_v^2}, \quad (12)$$

where G_h and G_v are the partial derivatives of the image intensity function in the horizontal and vertical directions, respectively.

Many FR-IQA indices follow a two-step procedure. First, a local quality map (LQM) is calculated by locally comparing the distorted image with the reference image through a similarity index. The LQM is a new image that captures the local quality of each zone of the distorted image. Finally, the quality index is calculated from the LQM image by a pooling strategy, which is a weighting of all of its pixel values. Several indices based on the magnitude of the gradient follow this type of scheme and use G to define an LQM. Two indices that follow this approach [21] are briefly reviewed below.

The gradient magnitude map of the reference image x is calculated as

$$Gx(i)=\sqrt{G_h(x)^2(i)+G_v(x)^2(i)} \quad (13)$$

where i denotes the location of pixel i in image x. Similarly, for the distorted image y,

$$Gy(i)=\sqrt{G_h(y)^2(i)+G_v(y)^2(i)}. \quad (14)$$

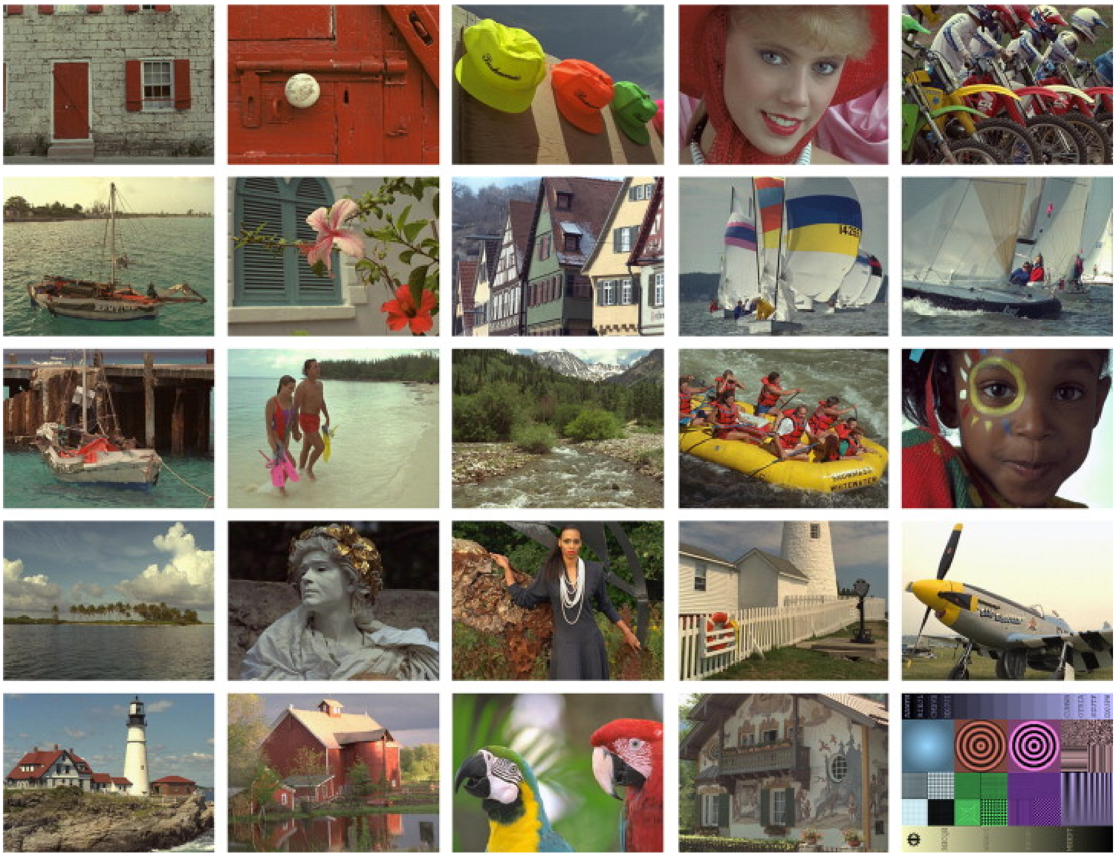


Figure. 1: Tampere Image Database 2013 (TID2013) reference images [5].

The gradient magnitude similarity (GMS) map is defined as a function of $Gx(i)$ and $Gy(i)$ through

$$\mathrm{GMS}(i) = \frac{2Gx(i)Gy(i) + c}{(Gx(i))^2 + (Gy(i))^2 + c}, \tag{15}$$

where $c > 0$ has been added to obtain numerical stability when the denominator of (15) is too small. Recall that the GMS is used as the LQM. Also note that if $x = y$, then $\mathrm{GMS}(i) = 1$, which is the maximum GMS value.

One of the simplest pooling strategies is to take the average of the local quality values as the final quality index. The *gradient magnitude similarity mean (GMSM)* follows this scheme and is hence defined as

$$\mathrm{GMSM} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{GMS}(i), \tag{16}$$

where N is the total number of pixels in the map. As a result, large GMSM values indicate high image quality levels.

The averaging process assumes that all pixels have the same importance in the image. However, in practice there are examples in which different regions contribute differently to the overall quality of an image. Because of this, there are proposals to assign a weight to local quality values before averaging them. This type of modification would give a more accurate result at the expense of highly increased computational costs [21].

When an image is distorted, local structures will suffer different degrees of degradation in the magnitude of the gradient. The *gradient magnitude similarity deviation (GMSD)* proposed in [21] notes that the spatial distribution of distortion levels has an impact on perception; that is, unevenly distributed distortion degrades visual quality more severely than do other types of distortion. GMSD uses the standard deviation as a pooling strategy to obtain the final quality index, because the global variation in local image quality is a reflection of its final quality. This index is defined as

$$\mathrm{GMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{GMS}(i) - \mathrm{GMSM})^2}. \tag{17}$$

The GMSD value accounts for the range of distortion severity of an image; therefore, the higher the GMSD value is, the greater the distortion range and thus the perceived quality. This index is both computationally efficient and effective in predicting quality [21].

6) *Visual information fidelity (VIF)*

The *visual information fidelity* [50] index is based on an information theory problem and is defined as the ratio between the joint information of images x and y and the marginal information of x through

$$\mathrm{VIF} = \frac{I(x, y + r)}{I(x, x + r)}, \tag{18}$$

where r is a stationary and Gaussian white noise process with variance σ_r^2. The denominator represents the information that the HVS can extract from the original image, while the numerator accounts for the information that the HVS can extract from the distorted image. Hence, (18) is highly correlated with the visual quality index studied in [15].

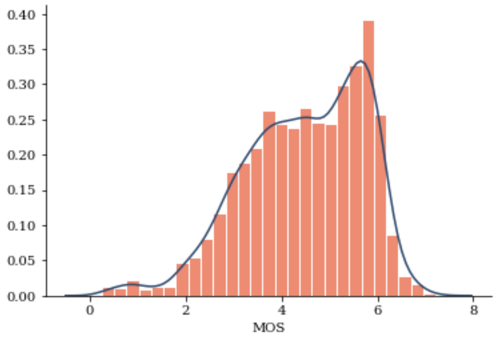


Figure. 2: MOS normalized histogram. The distribution is unknown and asymmetric.

C. *Image database*

Our research was sustained on the TID2013 image database [5], [6] as well as some FR-IQA indexes developed in recent years have been tested [3]. This database contains 3,000 distorted images obtained from 25 reference images (Figure 1), and each of these 3,000 images has an associated MOS value (Figure 2). In TID2013, 24 types of distortions (with 5 levels of distortion each) are considered. These 24 distortions are classified according to their characteristics [6] into the following groups: *noise*, *actual*, *simple*, *exotic*, *new* and *color* (as shown in Table 1). To obtain the quality of each of the 3,000 images according to each quality index, our own script routine is run generating a data frame.

To evaluate the normality of the set of 3,000 values that each index produced, Shapiro-Wilks and Kolmogorov-Smirnov tests were applied. Both tests showed p-values less than 0.0001 for all indices and for the MOS, indicating large deviations from normality in each case. As an example, the MOS distribution is displayed in Figure 2 and is clearly asymmetric. Instead of introducing transformation to achieve normality, in our analysis, nonparametric methods will be considered.

III. Proposed Method

A. *Comparative Analysis of Quality Indices using concordance coefficients*

Many authors propose to consider a quality index as successful if it is well correlated with the MOS index [6]. Following this, a correlation analysis between each index and the MOS was carried out, assuming that the MOS value of an image reflects its true quality. To avoid distributional assumptions, Spearman's and Kendall's rank correlation coefficients [51] were used.

Similar to other correlation coefficients, these coefficients vary between -1 and 1, and a coefficient equal to 0 implies that there is no correlation. Positive values close to 1 indicate that when one variable increases, so does the other, while negative values close to -1 indicate that when one variable increases, the other decreases. Correlations equal to -1 or 1 imply an exact monotonic relationship between the variables. Table 2 contains all possible interpretations. Spearman's and Kendall's rank correlation coefficients can also be used for testing the hypothesis of there being no correlation between two sequences [53].

To deepen the analysis, coefficients of concordance between the MOS and one or several quality indexes were included, considering that the correlation with the MOS is not enough to evaluate the performance of a quality index. While correlation attempts to quantify whether two datasets tend to vary in the same direction, concordance seeks to quantify whether two or more classifiers (in this case, the quality indices and the MOS) are equivalent. In general, concordance is a more restrictive concept than linear correlation, which measures the level of agreement between two variables by comparing the values with a specific straight line. Thus, a comparison between the MOS and a quality index based solely on the Spearman and Kendall coefficients is incomplete. In fact, concordance is a notion of agreement between the MOS and a given quality index that is not restricted to the linear correlation between them, hence it gives a more comprehensive understanding of the performance of a quality index.

In the next section, the concordance coefficients incorporated into the comparative analysis of quality indices are introduced.

1) *Kendall's coefficient of concordance*

Kendall's coefficient of concordance is a measure of agreement between m sets of n ranges. For instance, for a group of n objects evaluated by m judges, the coefficient provides information on the degree of agreement between the n ranges granted by the judges [17]. It is defined through

$$w = \frac{12S}{m^2(n^3-n) - m\sum_{j=1}^{m} L_j}, \tag{19}$$

where

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2,$$

R_i is the total rank given to object i, that is, the sum of all the ranks each judge j ($j \in \{1,...,m\}$) gave to object i ($i \in \{1,...,n\}$), and $\bar{R}$ is the mean value of those total ranges; finally, L_j is a correction factor for the set of ranks of judge j. If there are no ties, then $L_j = 0$.

Kendall's w always belong to the interval $[0,1]$. $w = 0$ may indicate that the attributes to be evaluated are ambiguous or poorly defined; then, there is no overall trend of agreement among judges, and their responses could be regarded as substantially random.

There is a hypothesis testing problem associated with (19), where the null hypotheses is H_0: Among the cited variables, there is no correlation. This test is a generalization of the Friedman test and is based on the previously defined w coefficient, a normalization of Friedman's statistic for the interval $[0,1]$.

2) *Cohen's kappa and Scott's pi concordance coefficients*

Cohen's kappa coefficient is another statistical measure of interrater agreement for qualitative items. It measures the agreement between the corresponding classifications of two evaluators who classify n elements into c mutually exclusive categories [18]. It is a more robust measure than simply a

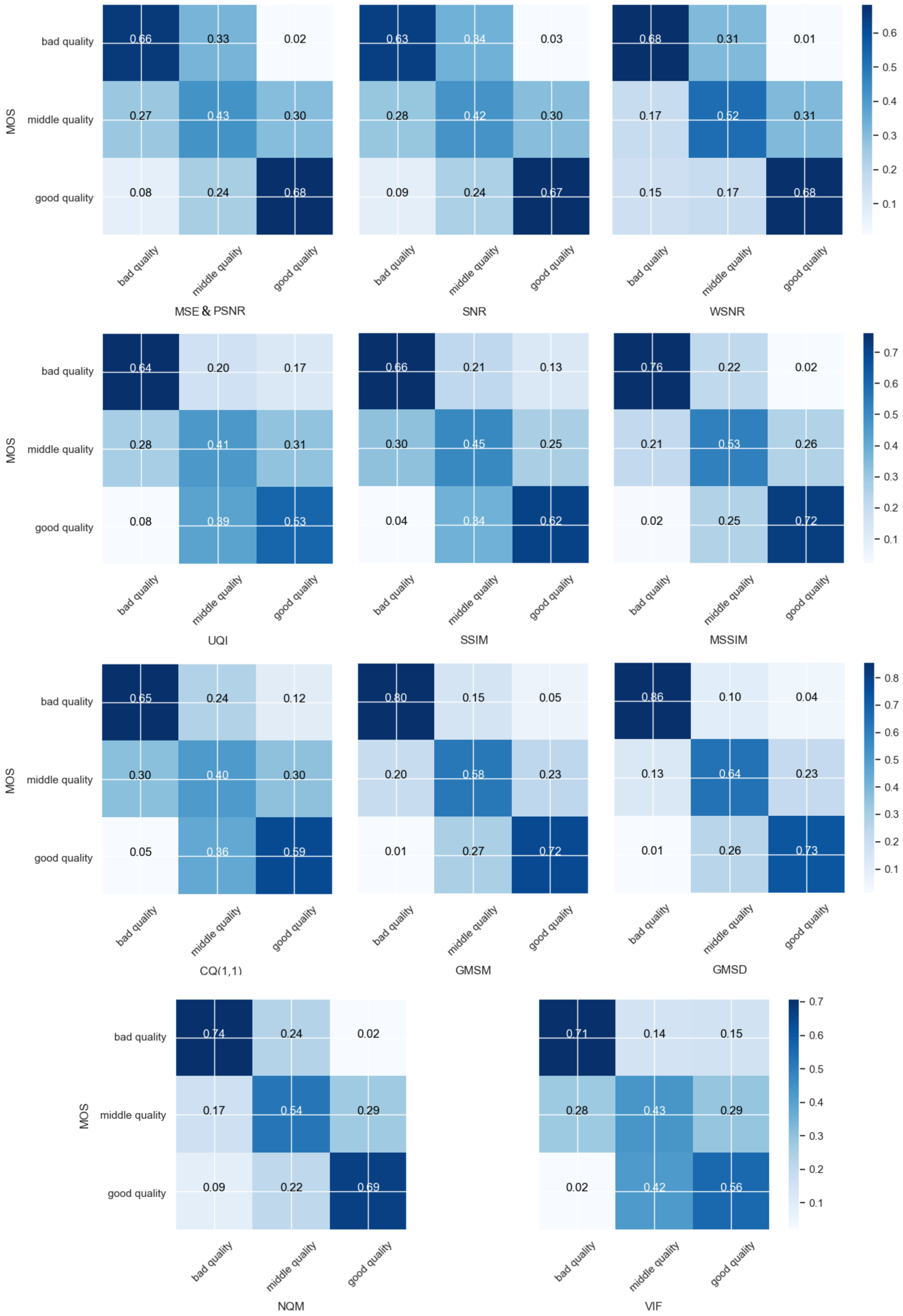


Figure. 3: Normalized confusion matrices for each quality index.

Table 1: Distortion types and their considered subsets present in TID2013 [6].

No.	Type of distortion	Noise	Actual	Simple	Exotic	New	Color	Full
1	Additive Gaussian noise	+	+	+	−	−	−	+
2	Additive noise in color comp.	+	−	−	−	−	+	+
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 transm. errors	−	−	−	+	−	−	+
13	JPEG2000 transm. errors	−	−	−	+	−	−	+
14	Non-ecc. pattern noise	−	−	−	+	−	−	+
15	Local blockwise dist.	−	−	−	+	−	−	+
16	Mean shift (intensity shift)	−	−	−	+	−	−	+
17	Contrast change	−	−	−	+	−	−	+
18	Change in color saturation	−	−	−	−	+	+	+
19	Multipl. Gaussian noise	+	+	−	−	+	−	+
20	Comfort noise	−	−	−	+	+	−	+
21	Lossy compr. of noisy images	+	+	−	−	+	−	+
22	Image color quant. w. dither	−	−	−	−	+	+	+
23	Chromatic aberrations	−	−	−	+	+	+	+
24	Sparse sampl. and reconstr.	−	−	−	+	+	−	+

Table 2: Interpretation of the estimated correlation coefficient [52].

Coefficient in the interval	Correlation
$(0; 0.2] \cup [-0.2; 0)$	Very Poor
$(0.2; 0.4] \cup [-0.4; 0 - .2)$	Poor
$(0.4; 0.6] \cup [-0.6; -0.4)$	Moderate
$(0.6; 0.8] \cup [-0.8; -0.6)$	Strong
$(0.8; 1] \cup [-1; -0.8)$	Very Strong

percentage agreement calculation because it takes into account the agreement that occurs randomly. It can also underestimate the agreement for a category of frequent use; because of this reason, it is considered a conservative measure. All these features render Cohen's kappa coefficient more robust for index comparison than those of Spearman and Kendall.

Scott's pi coefficient is similar to Cohen's kappa, differing only in the calculation of the probability of agreement by chance [54]. It assumes that the classifiers have the same distribution of responses, making Cohen's kappa slightly more informative. Scott's pi can be generalized to measure the concordance for several classifiers, leading to Fleiss' kappa coefficient, which allows for the obtaining of a multivariate concordance index between all the indices or a subset of them.

3) Fleiss' kappa coefficient

Fleiss' kappa is a generalization of Scott's pi for assessing the reliability of the agreement between a fixed number of reviewers by assigning categorical ratings to several items to be classified. The measure calculates the degree of agreement in the classification above what is expected by chance, assigning a score for the level of homogeneity or consensus there is between the scores given by the reviewers. It expresses how much the quantity observed in agreement between the reviewers exceeds what is expected from random qualifications [19].

To assess the agreement between the quality indices and the MOS, continuous variables were transformed into categorical variables. In [6], the authors classified the 3,000 TID2013 distorted images into three categories according to their MOS value: "bad quality", "middle quality" and "good quality". Following the same scheme, the 3,000 images were classified into these three categories according to each index, for which the one- and two-thirds percentiles delimited each category, and then, approximately 1,000 images were classified as being of "bad quality", 1,000 as being of "middle quality" and 1,000 as being of "good quality". In this fashion, it is possible to consider the quality indices and the MOS as classifiers and generate confusion matrices [55], in which the real classes are determined according with the MOS classification, while the predictions are determined by each index. As a result, Cohen's kappa and Scott's pi can be calculated from these confusion matrices, yielding a measure of the consistency of these qualifications.

B. Quality index ranking

Our goal is to determine which quality index has the best performance for each group of distortions and which ones follow in rank. If for a correlation (or concordance) coefficient, the quality indices are ordered from the one with the highest correlation (or concordance) with the MOS to the one with the lowest value, then a ranking of the quality indices is obtained. Because, in general, the coefficients take into account different factors, a number of dissimilar results could be obtained when using the four rankings generated according to the coefficients of Spearman, Kendall, Cohen and Scott. To overcome this inconvenience, the following system for the final quality index ranking is suggested: if m is the number of indices to be compared (in our case, $m = 12$), then for each ranking, the index that ranks first place adds $m - 1$ points, $m - 2$ for second place, $m - 3$ for third place, and so on.

Table 3: The first value shows the Spearman correlation between each quality index and the MOS index for each type of distortion's group. The second value (in bold) shows Kendall's correlation coefficients. At the bottom, the three new directional quality measures incorporated into the study are highlighted. All correlations obtained are significantly different from zero ($p < 0.0001$)

Index	Noise	Actual	Simple	Exotic	New	Color	Full
MSE	−0.7691 ; **-0.5619**	−0.7839 ; **-0.5762**	−0.8759 ; **-0.6892**	−0.5621 ; **-0.3923**	−0.7772 ; **-0.5760**	−0.7360 ; **-0.5373**	−0.6869 ; **-0.4958**
PSNR	0.7691 ; **0.5619**	0.7839 ; **0.5762**	0.8759 ; **0.6892**	0.5621 ; **0.3923**	0.7772 ; **0.5760**	0.7360 ; **0.5373**	0.6869 ; **0.4958**
SNR	0.7207 ; **0.5160**	0.7446 ; **0.5383**	0.8352 ; **0.6305**	0.5384 ; **0.3725**	0.7323 ; **0.5307**	0.6764 ; **0.4806**	0.6491 ; **0.4607**
WSNR	0.8711 ; **0.6827**	0.8691 ; **0.6813**	0.9227 ; **0.7551**	0.4298 ; **0.3046**	0.8872 ; **0.7054**	0.9014 ; **0.7210**	0.6382 ; **0.4938**
NQM	0.8482 ; **0.6557**	0.8507 ; **0.6598**	0.8882 ; **0.6997**	0.6116 ; **0.4336**	0.8762 ; **0.6924**	0.8937 ; **0.7080**	0.7126 ; **0.5348**
UQI	0.6030 ; **0.4194**	0.6403 ; **0.4494**	0.7348 ; **0.5230**	0.4718 ; **0.3311**	0.5191 ; **0.3650**	0.5092 ; **0.3613**	0.5239 ; **0.3695**
SSIM	0.6753 ; **0.4777**	0.7204 ; **0.5151**	0.7669 ; **0.5610**	0.5372 ; **0.3784**	0.7446 ; **0.5347**	0.6766 ; **0.4794**	0.6273 ; **0.4457**
MSSIM	0.8096 ; **0.6092**	0.8727 ; **0.6757**	0.8861 ; **0.6971**	0.7391 ; **0.5444**	0.7996 ; **0.5978**	0.7481 ; **0.5481**	0.7909 ; **0.5921**
VIF	0.7525 ; **0.5575**	0.8002 ; **0.6010**	0.8456 ; **0.6452**	0.5150 ; **0.3663**	0.7659 ; **0.5644**	0.7056 ; **0.5157**	0.6338 ; **0.4669**
CQ(1,1)	0.6509 ; **0.4657**	0.6531 ; **0.4691**	0.8356 ; **0.6358**	0.6458 ; **0.4590**	0.5284 ; **0.3790**	0.4821 ; **0.3444**	0.6009 ; **0.4292**
GMSM	0.8928 ; **0.7093**	0.8863 ; **0.7004**	0.9474 ; **0.7966**	0.7974 ; **0.6070**	0.6473 ; **0.5202**	0.6005 ; **0.4848**	0.7884 ; **0.6132**
GMSD	−0.9187 ; **-0.7461**	−0.9150 ; **-0.7408**	−0.9415 ; **-0.7949**	−0.8452 ; **-0.6528**	−0.6511 ; **-0.5248**	−0.5922 ; **-0.4723**	−0.8044 ; **-0.6339**

Table 4: Index ranking from highest to lowest correlation with the MOS according to the results of Table 3. Two indices are indicated when the Spearman and Kendall coefficients differ in their ordering.

	Noise	Actual	Simple	Exotic	New	Color	Full
1st	GMSD	GMSD	GMSM	GMSD	WSNR	WSNR	GMSD
2nd	GMSM	GMSM	GMSD	GMSM	NQM	NQM	MSSIM ; GMSM
3rd	WSNR	MSSIM ; WSNR	WSNR	MSSIM	MSSIM	MSSIM	GMSM ; MSSIM
4th	NQM	WSNR ; MSSIM	NQM	CQ(1,1)	MSE	MSE	NQM
5th	MSSIM	NQM	MSSIM	NQM	PSNR	PSNR	MSE
6th	MSE	VIF	MSE	MSE	VIF	VIF	PSNR
7th	PSNR	MSE	PSNR	PSNR	SSIM	SSIM ; GMSM	SNR ; WSNR
8th	VIF	PSNR	VIF	SNR ; SSIM	SNR	SNR	WSNR ; VIF
9th	SNR	SNR	CQ(1,1)	SSIM ; SNR	GMSD	GMSM ; SSIM	VIF ; SNR
10th	SSIM	SSIM	SNR	VIF	GMSM	GMSD	SSIM
11th	CQ(1,1)	CQ(1,1)	SSIM	UQI	CQ(1,1)	UQI	CQ(1,1)
12th	UQI	UQI	UQI	WSNR	UQI	CQ(1,1)	UQI

Then, the total summation score is obtained for each index. The index with the highest score will be ranked first, the one that follows will be ranked second, and so on.

This algorithm is explained with a pseudo-code and in Figures 4 and 5. Figure 4 illustrates a general scheme, while Figure 5 presents an application example. It should be considered that ties can occur, and thus, there may be no index that ranks in last place.

The scripts to generate the database with index values and classifications of the 3,000 images, the routines to calculate all the coefficients described in this section and the final ranking, as well as all the datasets and results, can be found online at `https://github.com/lucia15/IQA`.

Algorithm 1 Index ranking. Vector R contains the final ranking

```
A: set of distorted images
B: set of m indices
P: vector of length m
for index in B do
    for image in A do
        get image quality
    end for
    for $Spearman, Kendall, Cohen, Scott$ do
        get coeff qualities vs MOS
    end for
end for
$P = 0$
for $Spearman, Kendall, Cohen, Scott$ do
    $sort(B)$          ▷ sort indices according to each coeff
    for index in B do
        $r = position(index)$
        $P[index] += m - r$
    end for
end for
$R = argsort(P)$      ▷ get the indices of the sorted array
```

IV. Results

Table 3 shows the results of the Spearman and Kendall correlation analysis, suggesting that in the group of *simple* distortions, the quality indices are best correlated with MOS. Within the group of *exotic* distortions, it is observed that the lowest correlation with MOS for almost all quality indices except for the MSSIM and the directional indices: CQ(1,1), GMSM and GMSD. In addition, the last three indexes show a poor correlation with the MOS within the *color* and *new* distortion groups, respectively.

The confusion matrices between each quality index and the MOS are displayed in Figure 3. In all cases, the lowest agree-

Table 5: The first value shows Cohen's kappa concordance coefficient between the quality index and the MOS for each group of distortions; the second value (in bold) shows Scott's pi. These coefficients yielded the same results in the *full* group (the group containing all images).

Index	Noise	Actual	Simple	Exotic	New	Color	Full
MSE & PSNR	0.4334 ; **0.4309**	0.4502 ; **0.4463**	0.5223 ; **0.5188**	0.2826 ; **0.2728**	0.4702 ; **0.4681**	0.3944 ; **0.3943**	0.3820
SNR	0.3755 ; **0.3732**	0.3935 ; **0.3900**	0.4542 ; **0.4512**	0.2985 ; **0.2921**	0.4445 ; **0.4444**	0.3552 ; **0.3550**	0.3600
WSNR	0.4747 ; **0.4673**	0.4377 ; **0.4272**	0.5423 ; **0.5365**	0.3335 ; **0.3157**	0.5977 ; **0.5970**	0.6389 ; **0.6383**	0.4450
NQM	0.5120 ; **0.5084**	0.4907 ; **0.4849**	0.5454 ; **0.5425**	0.3919 ; **0.3833**	0.6043 ; **0.6036**	0.6429 ; **0.6427**	0.4845
UQI	0.2820 ; **0.2761**	0.3151 ; **0.3126**	0.4042 ; **0.4017**	0.2842 ; **0.2785**	0.2273 ; **0.2239**	0.2386 ; **0.2360**	0.2890
SSIM	0.3316 ; **0.3268**	0.3805 ; **0.3781**	0.4523 ; **0.4513**	0.3205 ; **0.3137**	0.3994 ; **0.3970**	0.3631 ; **0.3596**	0.3635
MSSIM	0.5102 ; **0.5093**	0.5826 ; **0.5813**	0.6446 ; **0.6437**	0.4365 ; **0.4340**	0.5050 ; **0.5039**	0.4270 ; **0.4223**	0.5095
VIF	0.3517 ; **0.3461**	0.4082 ; **0.4054**	0.4844 ; **0.4822**	0.2669 ; **0.2551**	0.3197 ; **0.3117**	0.2735 ; **0.2628**	0.3495
CQ(1,1)	0.3572 ; **0.3571**	0.3419 ; **0.3412**	0.5414 ; **0.5384**	0.2917 ; **0.2897**	0.2982 ; **0.2942**	0.3568 ; **0.3534**	0.3185
GMSM	0.6083 ; **0.6074**	0.5800 ; **0.5788**	0.7517 ; **0.7512**	0.4953 ; **0.4918**	0.4693 ; **0.4685**	0.5676 ; **0.5659**	0.5505
GMSD	0.7048 ; **0.7047**	0.6958 ; **0.6957**	0.7518 ; **0.7517**	0.5327 ; **0.5290**	0.5750 ; **0.5736**	0.5908 ; **0.5878**	0.6150

Table 6: Index ranking from highest to lowest concordance with the MOS according to the results of Table 5. Two indices are indicated when the Cohen and Scott coefficients differ in their ordering.

	Noise	Actual	Simple	Exotic	New	Color	Full
1st	GMSD	GMSD	GMSD	GMSD	NQM	NQM	GMSD
2nd	GMSM	MSSIM	GMSM	GMSM	WSNR	WSNR	GMSM
3rd	NQM ; MSSIM	GMSM	MSSIM	MSSIM	GMSD	GMSD	MSSIM
4th	MSSIM ; NQM	NQM	NQM	NQM	MSSIM	GMSM	NQM
5th	WSNR	MSE	WSNR ; CQ(1,1)	WSNR	MSE ; GMSM	MSSIM	WSNR
6th	MSE	PSNR	CQ(1,1) ; WSNR	SSIM	PSNR ; MSE	MSE	MSE
7th	PSNR	WSNR	MSE	SNR	GMSM ; PSNR	PSNR	PSNR
8th	SNR	VIF	PSNR	CQ(1,1)	SNR ; SNR	SSIM	SSIM
9th	CQ(1,1)	SNR	VIF	UQI	SSIM	CQ(1,1) ; SNR	SNR
10th	VIF	SSIM	SNR ; SSIM	MSE	VIF	SNR ; CQ(1,1)	VIF
11th	SSIM	CQ(1,1)	SSIM ; SNR	PSNR	CQ(1,1)	VIF	CQ(1,1)
12th	UQI	UQI	UQI	VIF	UQI	UQI	UQI

ment was for the "middle quality" category, as indicated by the blue scale.

Table 5 shows Cohen's kappa and Scott's pi coefficients calculated from the confusion matrices. The outcomes of these indices are very similar. The highest values were obtained mostly for the group of *simple* distortions. For *exotic* distortions, the values of the indices are the lowest in most of the cases. Broadly, Table 5 shows similar trends to those in Table 3, but with much lower values.

The quality indices were ordered from the highest to lowest values conforming to each correlation (or concordance) coefficient, obtaining four index rankings, which are shown in Tables 4 and 6. These rankings were pooled into a final ranking, as illustrated in Figure 5. Table 7 shows the final quality index ranking for each distortion group. For most of the distortion groups, the indices that performed best were the GMSD and the GMSM, except for the *new* and *color* distortion groups.

Figures 6 and 7 include a few scatterplots to check some of the obtained results. Figure 6 illustrates the main results shown in Table 7, while in Figure 7 some of the most common distortions in practice are examined: *additive Gaussian noise* (distortion #1), *Gaussian blur* (distortion #8), *JPEG compression* (distortion #10), and *JPEG2000 compression* (distortion #11).

Finally, Kendall's w concordance coefficient among the MOS and the 12 selected quality indices was computed, and the same was done for Fleiss' kappa. These two multivariate coefficients were also calculated among the MOS and the subset of indices that performed best in first, second, and third place for each distortion group, according to Table 7. It is observed that the agreement improves when the analysis is restricted to the subset of indices (Table 8).

V. Discussion

Faced with the problem of deciding which IQA measure to select to measure the similarity between two images, it is important to emphasize that there is no superior or optimal evaluation method. The selection of one or another tool is often done in an arbitrary way, without considering the type of distortion that could be affecting the images to be compared, nor the limitations of the comparison index. In this work, an attempt was made to address this problem and provide the researchers with a methodology that helps them make a decision based on statistical foundations that incorporate the agreement between the methods, in addition to the correlation. A method that provides a guide for comparing the performance of any set of FR-IQA indices given different types of distortions was described. The method has been used to carry out a comprehensive comparative analysis of a set of FR-IQA indices under a variety of distortions. To the best of our knowledge, this is the first attempt that incorporates concordance measures to the study of FR-IQA performance.

A. Conclusion

Concordance measures revealed aspects that have not been taken into account by traditional correlation coefficients. Co-

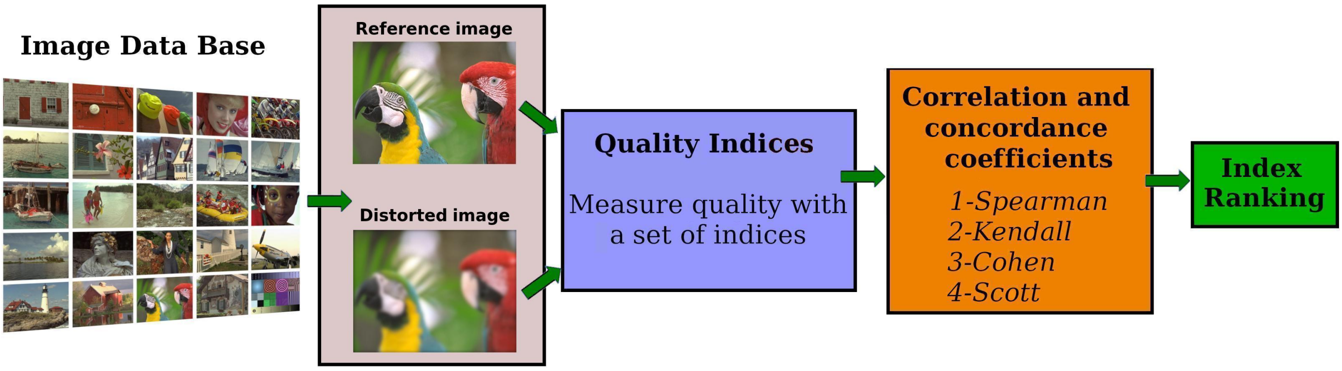
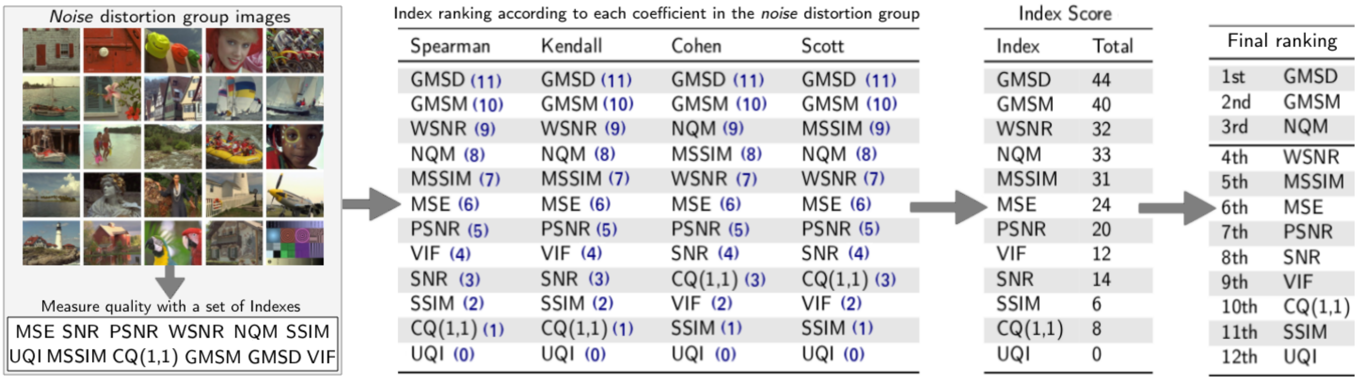


Figure. 4: Work flow to obtain the quality index ranking. The correlation and concordance coefficients are calculated between each quality index and the MOS.


Noise distortion group images
Measure quality with a set of Indexes
MSE SNR PSNR WSNR NQM SSIM
UQI MSSIM CQ(1,1) GMSM GMSD VIF

Index ranking according to each coefficient in the *noise* distortion group

Spearman	Kendall	Cohen	Scott
GMSD (11)	GMSD (11)	GMSD (11)	GMSD (11)
GMSM (10)	GMSM (10)	GMSM (10)	GMSM (10)
WSNR (9)	WSNR (9)	NQM (9)	MSSIM (9)
NQM (8)	NQM (8)	MSSIM (8)	NQM (8)
MSSIM (7)	MSSIM (7)	WSNR (7)	WSNR (7)
MSE (6)	MSE (6)	MSE (6)	MSE (6)
PSNR (5)	PSNR (5)	PSNR (5)	PSNR (5)
VIF (4)	VIF (4)	SNR (4)	SNR (4)
SNR (3)	SNR (3)	CQ(1,1) (3)	CQ(1,1) (3)
SSIM (2)	SSIM (2)	VIF (2)	VIF (2)
CQ(1,1) (1)	CQ(1,1) (1)	SSIM (1)	SSIM (1)
UQI (0)	UQI (0)	UQI (0)	UQI (0)

Index Score

Index	Total
GMSD	44
GMSM	40
WSNR	32
NQM	33
MSSIM	31
MSE	24
PSNR	20
VIF	12
SNR	14
SSIM	6
CQ(1,1)	8
UQI	0

Final ranking	
1st	GMSD
2nd	GMSM
3rd	NQM
4th	WSNR
5th	MSSIM
6th	MSE
7th	PSNR
8th	SNR
9th	VIF
10th	CQ(1,1)
11th	SSIM
12th	UQI

Figure. 5: Steps to obtain the quality index ranking: *noise* distortion group example. Step 1. Measure quality of each distorted image according to each quality index. Step 2. Calculate Spearman, Kendall, Cohen and Scott coefficients between each quality index values and MOS values. Step 4. Sort them from highest to lowest. Step 4. Assign 11 points to the first index, 10 to the second, and so on. Step 5. Calculate the total that adds up each index and order them from highest to lowest to establish the final ranking.

hen's and Scott's indices have shown similar trends as those of Spearman and Kendall, but for lower values, they are less optimistic about the degree of agreement with the MOS. One reason for this could finding be the robustness of these measures. In addition, these measures are also more realistic, although in practice, most of the quality indices do not attain the desired performance.

Our experiments reveal that the FR-IQA indices are sensitive to the type of distortions. In the group of *simple* distortions, the quality indices were quite in agreement with the MOS, while the group of *exotic* distortions had the worst performance. It is remarkable that for the *new* and *color* distortion groups, the most appropriate indices were the WSNR and NQM, unlike the other groups, in which it was preferable to use a directional index. In the case of image compression (distortions #10 and #11), *additive Gaussian noise* (distortion #1) and *Gaussian blur* (distortion #8), the GMSD was proven to be the most appropriate index.

In practice, it is difficult to know the type of distortion that affects an image. Additional information such as the transmission or capturing method, is commonly needed. For instance, if an image is sent by mail, then it will be affected by compression, as well as possibly by other distortions. Although in practice, many times, the actual type of distortion is unknown, this work reveals that even ignoring this factor, the indices that showed the best performance are the GMSD, GMSM and MSSIM, in decreasing order.

An interesting contribution of this work is that it expanded the study of Ponomarenko et al. by incorporating the directional indices of the GMSD, GMSM and CQ(1,1) into the correlation tables proposed in [6]. An important finding is that precisely two of these new indices best correlate with the MOS both in general and per distortion group, except in the *new* and *color* groups. Furthermore, new tables based on concordance coefficients were proposed. The combination of all these strategies led to a more complete comparison method, from which a ranking of quality indices could be generated from any set of them.

All the scripts, generated databases, and results are available at `https://github.com/lucia15/IQA`, hoping that others might benefit from open-source implementations organized in a single repository. This is an extra contribution in an area in which reproducible research is often difficult.

B. Future Work

Although the directional indices—the GMSD and GMSM—showed the highest correlation with the MOS in the *full* group,

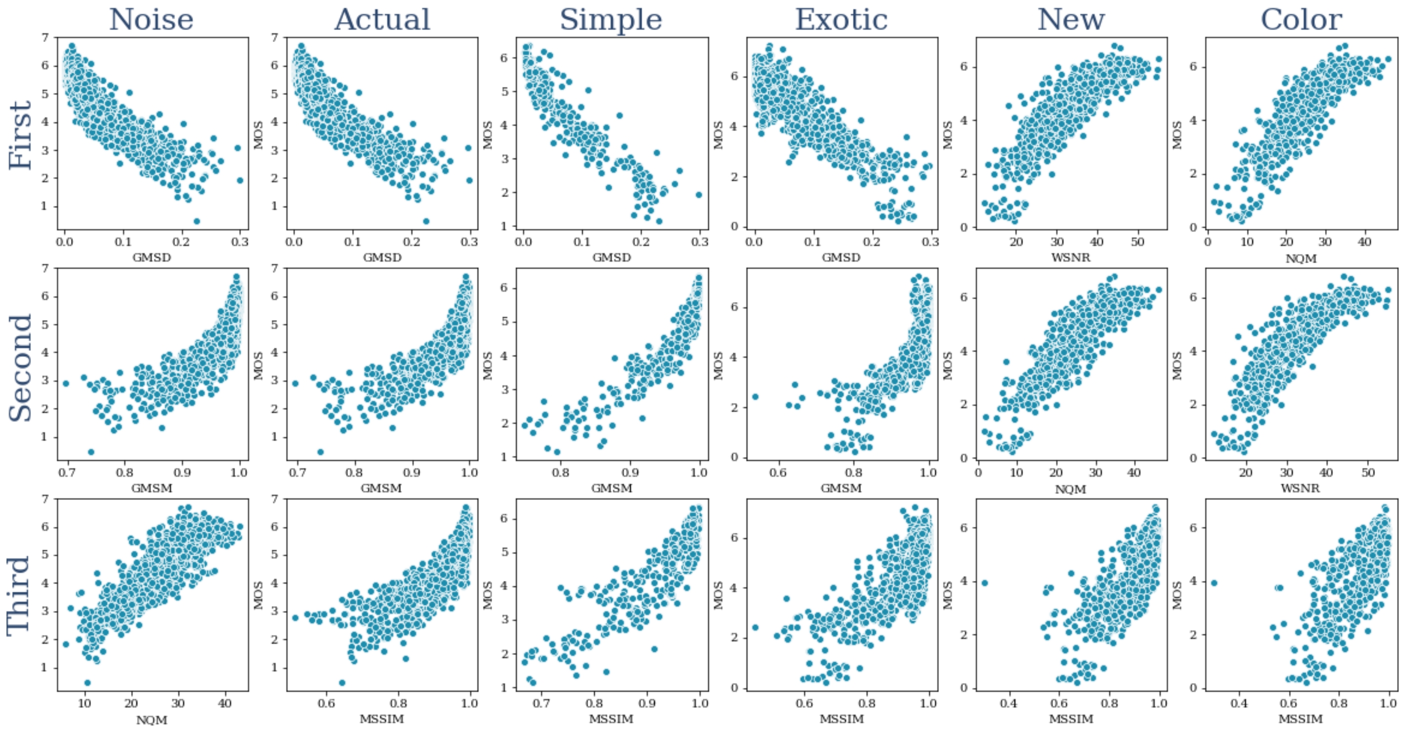


Figure. 6: Scatterplots showing the correlation between MOS and the quality indices with the best performance for each group of distortions.

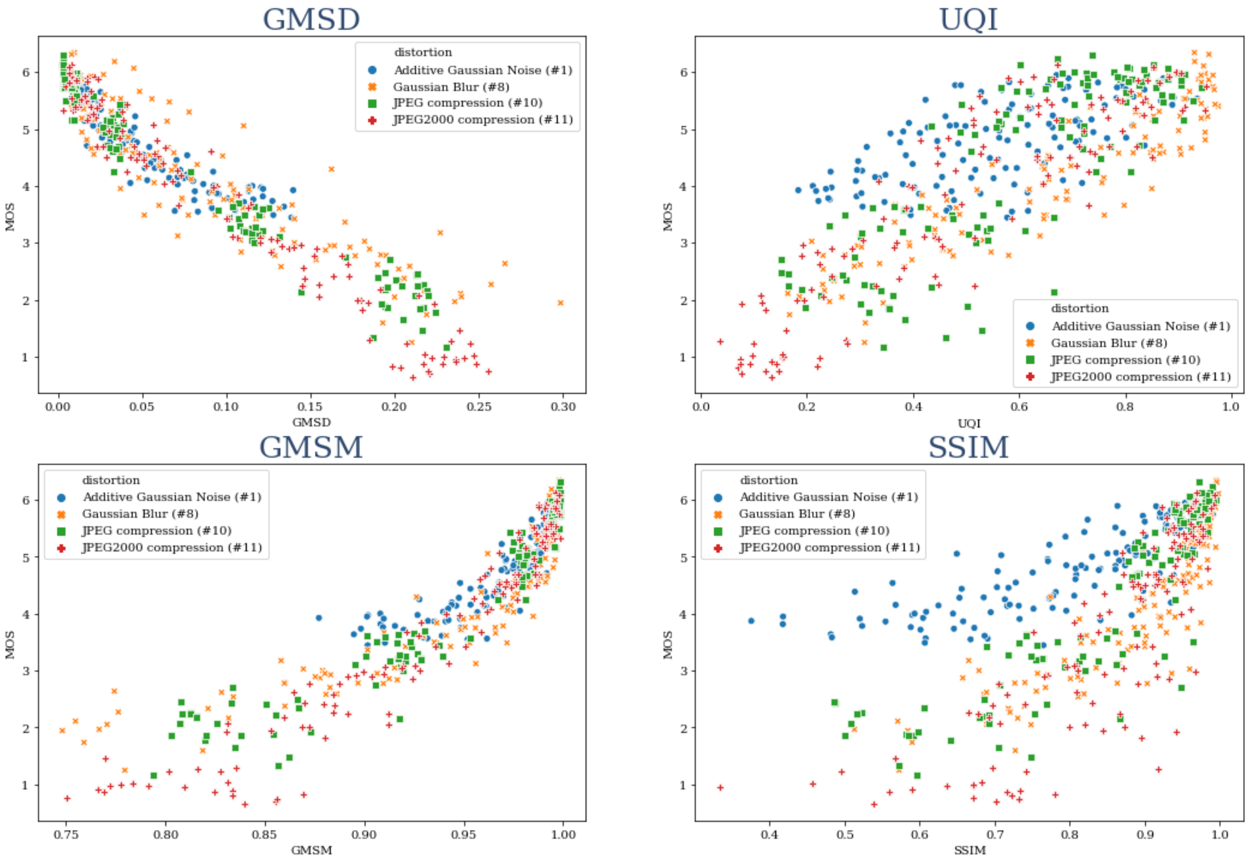


Figure. 7: Scatterplots for the most common distortions. The plots on the left show a strong correlation, while those on the right show a poor correlation.

Table 7: Most appropriate quality index per distortion group.

	Noise	Actual	Simple	Exotic	New	Color	Full
1st	GMSD	GMSD	GMSD-GMSM	GMSD	NQM-WSNR	NQM-WSNR	GMSD
2nd	GMSM	GMSM	MSSIM-NQM	GMSM	MSSIM	MSSIM	GMSM
3rd	NQM	MSSIM	WSNR	MSSIM	MSE	MSE	MSSIM
4th	WSNR	NQM	MSE	NQM	PSNR	GMSM-PSNR	NQM
5th	MSSIM	WSNR	CQ(1,1)	CQ(1,1)	GMSD	GMSD	MSE
6th	MSE	MSE	PSNR	SSIM	GMSM-SNR	SSIM	WSNR
7th	PSNR	PSNR-VIF	VIF	SNR	SSIM-VIF	VIF	PSNR
8th	SNR	SNR	SNR	MSE	CQ(1,1)	SNR	SNR
9th	VIF	SSIM	SSIM	WSNR	UQI	CQ(1,1)	SSIM
10th	CQ(1,1)	CQ(1,1)	UQI	PSNR	—	UQI	VIF
11th	SSIM	UQI	—	UQI	—	—	CQ(1,1)
12th	UQI	—	—	VIF	—	—	UQI

Table 8: Kendall's w and Fleiss' κ among the MOS and the quality indices by distortion group. In each column, the first values correspond to the coefficient among the MOS and all quality indices, whereas the second values in bold correspond to the coefficient among the MOS and the subset of indices that worked best in first, second, and third place for each distortion group, according to Table 7. All w values are significantly different from zero ($p < 0.0001$) according to Kendall's concordance test.

Group	Kendall's w	Fleiss' κ
Noise	0.7802 ; **0.9357**	0.4329 ; **0.5896**
Actual	0.7986 ; **0.9418**	0.4518 ; **0.6719**
Simple	0.8544 ; **0.9358**	0.5134 ; **0.6471**
Exotic	0.5894 ; **0.8726**	0.3251 ; **0.5084**
New	0.7472 ; **0.8200**	0.4685 ; **0.5561**
Color	0.7043 ; **0.7855**	0.4280 ; **0.5339**
Full	0.6659 ; **0.8658**	0.4020 ; **0.5913**

no matter the type of distortion used, this does not imply that the distortions have any directional factor but rather that the indices based on the magnitude of the gradient captured the similarity between images more often than other indices did. Therefore, it is proposed to elucidate in future research what makes this type of index so effective in capturing similarity. In the case of the CQ index, it has only been considered the direction $h = (1, 1)$. It is planned to approach the problem of how to choose a set of directions of interest to obtain a summary function of these CQ values in several directions, e.g., along the lines given in [56].

One problem that is related to the methodology suggested in this paper is the inclusion of the spatial concordance correlation coefficient (SCCC) recently studied in [25]. This coefficient is a generalization of the concordance index introduced in [57] and has two main features. First, it preserves the interpretation of Lin's index in the sense that it evaluates the agreement between two continuous variables by measuring their joint deviation from a 45° line through the origin. Second, it measures the spatial concordance between two georeferenced variables for a fixed value of the spatial lag h. The exploration of the SCCC in the context of image quality assessment remains an open problem that should be addressed in future research.

Another pending task is to determine desirable characteristics for a quality index. In this direction, we propose to continue the ideas raised in [35].

Acknowledgments

This work was supported by *Secretaría de Ciencia y Tecnología* (SeCyT) from *Universidad Nacional de Córdoba* (UNC) Proyecto Consolidar 2018-2121 (P.I.D. Nº 33620180100055CB) and *Consejo Nacional de Investigaciones Científicas y Técnicas* (CIEM - CONICET). R. Vallejos acknowledges financial support from CONICYT through the MATH-AMSUD program, grant 20-MATH-03, from UTFSM, grant P_LIR_2020_20, and from the Advanced Center for Electrical and Electronic Engineering (AC3E), grant FB-0008. The authors would like to give special thanks to Álvaro Cuestas, Mauro Pappaterra and Francisco Trucco for their writing of revisions. [26] [12] [27][28][11]

References

[1] Z. Wang, A. Bovik, Modern image quality assessment, Synthesis Lectures on Image, Video, and Multimedia Processing 1 (2006) 1–156.

[2] T. Pappas, R. Safranek, Percentual criteria for image quality evaluation, A. C. Bovik, Ed. New-York: Academic, 2000.

[3] C. Deng, L. Ma, W. Lin, K. Ngan, Visual Signal Quality Assessment, Springer International Publishing, 2015.

[4] N. Ponomarenko, V. Lukin, A. Zelensky, A. Egiazarian, M. Carli, F. Battisti, Tid2008- a database for evaluation of full-reference visual quality assessment metrics, Advances of Modern Radioelectronics 10 (2009) 30–45.

[5] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C. Jay Kuo, Color image database tid2013: Peculiarities and preliminary results, 4th European Workshop on Visual Information Processing EUVIP2013 (2013) 106–111.

[6] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C. Jay Kuo, Image database tid2013: Peculiarities, results and perspectives, Signal Processing: Image Communication 30 (2014) 57–77.

[7] E. Larson, D. Chandler, Most aparent distortion: Full reference image quality assessment and the rol of the strategy, Journal of Electronic Imaging 19 (2010) 011006.

[8] D. Narsaiah, R. S. Reddy, A. Kokkula, P. A. Kumar, A. Karthik, A novel full reference-image quality assessment (fr-iqa) for adaptive visual perception improvement, in: 2021 6th International Conference on Inventive Computation Technologies (ICICT), IEEE, 2021, pp. 726–730.

[9] N. Abbas, T. Saba, S. Khan, Z. Mehmood, A. Rehman, R. Tabasum, Reduced reference image quality assessment technique based on dwt and path integral local binary patterns, Arabian Journal for Science and Engineering 45 (2020) 3387–3401.

[10] L. Wei, L. Zhao, J. Peng, Reduced reference quality assessment for image retargeting by earth mover's distance, Applied Sciences 11 (2021) 9776.

[11] L. Li, T. Song, J. Wu, W. Dong, J. Qian, G. Shi, Blind image quality index for authentic distortions with local and global deep feature aggregation, IEEE Transactions on Circuits and Systems for Video Technology (2021).

[12] H. Khalid, M. Ali, N. Ahmed, Gaussian process-based feature-enriched blind image quality assessment, Journal of Visual Communication and Image Representation 77 (2021) 103092.

[13] Y. Ding, Visual Quality Assessment for Natural and Medical Image, Springer Berlin Heidelberg, 2018.

[14] Z. Wang, A. Bovik, Mean squared error: love it or leave it? - a new look at fidelity measures, IEEE Signal Process, Magazine 26 (2009) 98–117.

[15] H. Sheikh, M. Sabir, B. A.C., A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Processing 15 (2006) 3449–3451.

[16] K. Silpa, S. Mastani, Comparison of image quality metrics, International Journal of Engineering Research & Technology 1 (2012).

[17] M. G. Kendall, S. B. Babington, The problem of m rankings, The Annals of Mathematical Statistics 10 (1939) 275–287.

[18] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37–46.

[19] J. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (1971) 378–382.

[20] S. Ojeda, R. Vallejos, P. Lamberti, Measure of similarity between images based on the codispersion coefficient, Journal of Electronic Imaging 21 (2012) 023019.

[21] W. Xue, L. Zhang, X. Mou, A. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE Trans. Image Processing 23 (2014) 684–695.

[22] S. Ojeda, G. Britos, R. Vallejos, An image quality index based on coefficients of spatial association with an application to image fusion, Spatial Statistics 23 (2018) 1–16.

[23] W. Lin, J. Kuo, Perceptual visual quality metrics: A survey, J. Vis. Commun. Image R 22 (2011) 297–312.

[24] S. Pistonessi, J. Martinez, S. Ojeda, R. Vallejos, Structural similarity metrics for quality image fusion assessment: Algorithms, Image Processing On Line 8 (2018) 345–368.

[25] R. Vallejos, J. , Pérez, A. , Ellison, A. Richardson, A spatial concordance correlation coefficient with an application to image analysis, Spatial Statistics (2020 in press).

[26] M. Frackiewicz, G. Szolc, H. Palus, An improved spsim index for image quality assessment, Symmetry 13 (2021) 518.

[27] C. Shi, Y. Lin, Full reference image quality assessment based on visual salience with color appearance and gradient similarity, IEEE Access 8 (2020) 97310–97320.

[28] L. Wang, D. Rajan, An image similarity descriptor for classification tasks, Journal of Visual Communication and Image Representation 71 (2020) 102847.

[29] L. Wang, A survey on iqa, arXiv preprint arXiv:2109.00347 (2021).

[30] F. Osorio, R. Vallejos, W. Barraza, S. Ojeda, M. Landi, Statistical estimation of the structural similarity index for image quality assessment, Signal, Image and Video Processing (2021) 10.1007/s11760–021–02051–9.

[31] K. Ding, K. Ma, S. Wang, E. P. Simoncelli, Comparison of full-reference image quality models for optimization of image processing systems, International Journal of Computer Vision 129 (2021) 1258–1281.

[32] M. Yang, G. Yin, Y. Du, Z. Wei, Pair comparison based progressive subjective quality ranking for underwater images, Signal Processing: Image Communication 99 (2021) 116444.

[33] K. Ding, K. Ma, S. Wang, E. P. Simoncelli, Image quality assessment: Unifying structure and texture similarity, arXiv preprint arXiv:2004.07728 (2020).

[34] K. Egiazarian, M. Ponomarenko, V. Lukin, O. Ieremeiev, Statistical evaluation of visual quality metrics for image denoising, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 6752–6756.

[35] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, L. Zhang, Waterloo exploration database: New challenges for image quality assessment models, IEEE Transactions on Image Processing 26 (2016) 1004–1016.

[36] T. Mitsa, K. Varkur, Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms, IEEE International Conference on Acoustics, Speech, and Signal Processing 5 (1993) 301–304.

[37] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, A. Bovik, Image quality assessment based on a degradation model, IEEE Trans. on Image Processing 9 (2000) 636–650.

[38] Z. Wang, A. Bovik, A universal image quality index, IEEE Signal Processing Letters 9 (2002) 81–84.

[39] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Processing 13 (2004) 600–612.

[40] Z. Wang, E. Simoncelli, A. Bovik, Multi-scale structural similarity for image quality assessment, IEEE Asilomar Conference Signals, Systems and Computers 2 (2003) 1398–1402.

[41] D. Chandler, S. Hemami, Vsnr: A wevelet-based visual sinal-to-noise-ratio for natural images., IEEE Trans. Image Process. 9 (2007) 2284–2298.

[42] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Trans. Image Processing 20 (2011) 1185–1198.

[43] L. Zhang, L. Zhang, X. Mou, D. Zhang, Fsim: A feature similarity index for image quality assessment, IEEE Trans. Image Processing 20 (2011) 2378–2386.

[44] S. Wang, A. Rehman, Z. Wang, S. Ma, W. Gao, Ssim-motivated rate-distortion optimization for video coding, IEEE Transactions on Circuits and Systems for Video Technology 22 (2012) 516–529.

[45] K. Ma, K. Seng, Z. Wang, Perceptual quality assessment for multi-exposure image fusion, IEEE Transactions on Image Processing 24 (2015) 3345–3356.

[46] W. Hans, Multivariate Geostatistics: An Introduction with Appplications, Springer Berlin Heidelberg, 2003.

[47] N. Cressie, Statistics for Spatial Data, Wiley, 1993.

[48] A. Rukhin, R. Vallejos, Codispersion coefficient for spatial and temporal series, Statistics & Probability Letters 78 (2008) 1290–1300.

[49] R. Vallejos, D. Mancilla, J. Acosta, Image similarity assessment based on coefficients of spatial association, Journal of Mathematical Imaging and Vision 56 (2016) 77–98.

[50] H. Sheikh, A. Bovik, Image information and visual quality, IEEE Trans. Image Processing 15 (2006) 430–444.

[51] C. Spearman, The proof and measurement of association between two things, The American Journal of Psychology 15 (1904) 72–101.

[52] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 1 (2006) 1–156.

[53] G. W. Corder, D. I. Foreman, Nonparametric Statistics, Wiley, 2014.

[54] S. W. A., Reliability of content analysis: The case of nominal scale coding, The Public Opinion Quarterly 19 (1955) 321–325.

[55] A. Tharwat, Classification assessment methods, Applied Computing and Informatics (2018 in press).

[56] R. Vallejos, A. Mallea, M. Herrera, S. Ojeda, A multivariate geostatistical approach for landscape classification from remotely sensed image data, Stochastic Environmental Research and Risk Assessment 29 (2015) 369–365.

[57] L. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1989) 255–268.

Author Biographies

María Lucía Pappaterra received her Bachelor's degree in Mathematics Teaching from *Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba* in March 2011. Since April 2017, she has been working as a Ph.D. student in the same institution at the Department of Probability and Statistics. In December 2018 she received a Master's degree in Applied Statistics. Her current research interests include image analysis, visual perception, and signal processing.

Silvia María Ojeda received her PhD in mathematics from the *Universidad Nacional de Córdoba, Argentina* in 1999. She joined the *Group of Probability and Statistics* at *Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba* in 2003. Her current research topics include spatial statistics, robust modeling, time series and statistical image processing.

Marcos Alejandro Landi received his Bachelor's degree in Biology from *Facultad de Ciencias Exactas, Físicas y Naturales* (FCEFyN), *Universidad Nacional de Córdoba* (UNC), Argentina in 2008; and the Ph. D. Degree in Biological Sciences from the same institution in

2018. Since 2018, he has been working at the *Instituto de Diversidad y Ecología Animal* (IDEA-CONICET) as a remote sensing and GIS specialist. His research interests include the development of statistical methods to analyze time series obtained from remote sensing, landscape fire ecology, and methodological developments to analyze and process images.

Ronny Obed Vallejos received his Bachelor and Master of Science in Mathematics from the Universidad Técnica Federico Santa María, Chile in 1995 and 1998, respectively. He also received his Master degree in Statistics from the University of Connecticut, USA in 2002, and later his Ph.D. in Statistics from the University of Maryland Baltimore County in 2006. Currently, he is an associate professor in the Department of Mathematics at the Universidad Técnica Federico Santa María, Chile. His research interests are spatial statistics, statistical image processing, time series, and agreement measures.