J. Vis. Commun. Image R. 90 (2023) 103704

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Proposal of a new fidelity measure between computed image quality and observers quality scores accounting for scores variability ☆

Pedro Latorre-Carmona [a,*], Rafael Huertas [b], Marius Pedersen [c], Samuel Morillas [d]

[a] Departamento de Ingeniería Informática, Universidad de Burgos, Spain
[b] Departamento de Óptica, Universidad de Granada, Spain
[c] Colourlab, Department of Computer Science, Norwegian University of Science and Technology, Norway
[d] Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Spain

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
STRESS
Psycophysics
Image quality metric
Evaluation

ABSTRACT

Assessment of the visual quality of colour images is usually a difficult process, validated through hard-to-carry-out psychophysical experiments, used to record observer quality scores. Visual image quality metrics aim to maximise the agreement between computed indexes and observer scores, or opinions. Therefore, in this area, it is of critical importance to have appropriate measures of this agreement (i.e. performance) between the computed image quality metric values and observer's quality scores, both for the development, as well as for the use of image quality metrics. Among the measures of agreement, the most used one nowadays is the well-known Pearson correlation coefficient, while Spearman rank correlation coefficient is also commonly used. The aim of this paper is two-fold. First, to introduce the Standardised Residual Sum of Squares ($STRESS$) as an alternative metric for the agreement between computed image quality and observers quality scores and analyse its properties and advantages in front of Pearson, Spearman and Kendall correlation coefficients; Second, to introduce a new version of $STRESS$ (called $USTRESS$) that takes observers' scores variability into account. The results on synthetic and real datasets support that $STRESS$ has a series of benefits in front of the classical approaches and that the inclusion of uncertainty in $STRESS$ has an important effect on the results, quantified by statistical significance tests. A free to download MATLAB code version of $USTRESS$ is available at https://viplab.webs.upv.es/resources/

1. Introduction

Images and multimedia information are a type of data that is increasingly present in our lives. The assessment of their quality is an active research topic nowadays in the areas of computer vision, entertainment, colour vision, and perception, just to cite a few of them.

Image quality can be inferred by subjective and by objective methods. The former ones are based on the perceptual assessment (or the opinion, or a numerical assignment) of a human viewer about an (or a group of) attribute(s) of an image or a set of them. Objective methods use computational models that somehow assign a *perceptual* quality value to the images. These are commonly referred to as Image quality Metrics (IQMs). IQMs can be divided into three main types depending on the availability of the reference image [1]: full-reference, where the reference image (original) and test image (distorted) are used to estimate perceptual quality; reduced-reference, where partial information of the reference and test images are used; and no-reference, where only the test image is used.

Inferring the visual quality of colour images, which needs to be done through psychophysical experiments used to record observers' scores, is a costly and hard process. This is the reason why it is important to have IQMs available that might avoid these experiments. Obviously, the aim of any IQMs is to maximise the agreement with observers scores which are considered the *true* value.

Among the full-reference IQMs used so far [1], one of the most relevant currently used ones is the structural similarity index (SSIM) [2], which represents more than 23 K citations according to Scopus [3] nowadays. Other IQMs are: (a) the so-called image Color Appearance Model difference (iCAM) [4]; (b) The Fuzzy Color Structural Similarity (FCCS) [5]; (c) The Multiscale version of SSIM (MSSIM) [6]; (d) The Color Structural Similarity Index (CSSIM) [7]; (e) The Feature Similarity Index (FSIMc) [8]; (f) The Mean Squared Error (MSE); (g) The

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: plcarmona@ubu.es (P. Latorre-Carmona), rhuertas@ugr.es (R. Huertas), marius.pedersen@ntnu.no (M. Pedersen), smorillas@mat.upv.es (S. Morillas).

https://doi.org/10.1016/j.jvcir.2022.103704
Received 4 July 2022; Received in revised form 28 October 2022; Accepted 19 November 2022
Available online 30 November 2022

Root-Mean-Square Error (RMSE); (h) The Peak Signal-To-Noise Ratio (PSNR); and (i) The Normalised Color Difference (NCD) [9]. In total, we have considered 10 of the most representative IQMs currently in use. iCAM is a difference measure derived from the iCAM image Color Appearance Model that is in turn an extension of a Color Appearance Model (CAM) to include aspects concerning spatial effects on the perception of colours. SSIM uses three different measures of similarity (luminance, contrast, and structure) between image patches that are combined to obtain a single similarity degree. FCSS is an extension to the fuzzy context of SSIM where highly nonlinear fuzzy measures are used to measure the luminance, contrast and structure similarities. MSSIM is similar to SSIM but applied over multiple spatial scales, by considering a multi-stage sub-sampling process, in a similar way as that considered is made in the early vision system. CMSSIM is an extension of the MSSIM merit figure, where the contrast and structure terms of SSIM are applied at each scale, the luminance comparison is applied at the highest scale only, and the colour comparison is applied only at the lowest scale. FSIMc is based on the fact that the human visual system processes and analyses an image taking into account mainly its low-level features. In particular, FSIMc considers phase congruency (a dimensionless measure of the significance of a local structure) and the image gradient magnitude as the two features to obtain a local quality map and a final image quality score. Other classical measures from the signal processing field include: MSE (measures the average squared difference between the estimated values and the actual values to be compared against), RMSE (represents the quadratic mean of the differences between the predicted and the corresponding observed values), PSNR computes the ratio between the maximum possible value of a signal (variable) and the power of corrupting noise that affects it (obtained from MSE) and it is usually expressed in logarithmic units. Finally, NCD works by comparing all pairs of colour pixels in a more perceptually uniform colour space (for instance, the CIELUV colour space) instead of the RGB space, computing their Euclidean distances, and normalising by the norm of one of the two vectors.

There also exist a wide range of no-reference IQMs, which only use the test (distorted) image for estimating image quality. These include the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [10] that is a natural scene statistic-based distortion-generic IQM in the spatial domain. The Natural Image Quality Evaluator (NIQE) [11] uses a set of statistical features from natural scene statistic models. The Perception-based Image Quality Evaluator (PIQE) [12] is based on how human perceive images and exploits local features. The blurMetric [13] is based on comparing the discrimination between different levels of blur, while CEIQ [14] is based on calculating the degree of deviation from the natural scene statistics models. The Cumulative Probability of Blur Detection (CPBD) [15] uses a probabilistic model to calculate the probability of being able to detect blur at edges in images. Feature maps based Referenceless Image QUality Evaluation Engine (FRIQUEE) [16] uses a bag of feature-maps approach, which does not make assumptions on the distortions present in images. The Just-Noticeable-Blur Measure (JNBM) [17] uses the notion of just-noticeable-blur to estimate sharpness. In addition to these there exist a vast literature on no-reference IQMs.

Within this context, with the wide proposal of IQMs, it is easy to see that it is necessary and of critical importance to measure the agreement between image visual and computed quality in a solid and reliable way. This is done to assess the performance of the IQMs in terms of finding how well an IQM is able to predict perceived image quality.

For this task, the most used indicators are the well-known Pearson correlation coefficient [18], the Spearman rank correlation coefficient [19], and Kendalls correlation [20]. The Pearson correlation coefficient measures the linear relationship between the IQM values and the subjective scores. A higher correlation between the IQM values and subjective scores indicates higher performance. On the other hand, the Spearman and Kendalls correlation coefficients measure the relationship between the IQM rankings and the subjective scores rankings, using a monotonic function. Often a non-linear regression is applied to deal with the fact that the relationship between subjective scores and metric values can be of non-linear nature [21]. More recently, the Perceptually Weighted Rank Correlation (PWRC) [22] was proposed a performance measure that awards correct ranking of high-quality images and at the same time reduces the importance of mistakes in insensitive ranking. I still miss a reference to Kendal

An alternative strategy that we analyse in this paper is the $STRESS$ metric, which was originally used in multidimensional scaling (MDS) techniques [23,24], and it has been extensively used afterwards to measure the agreement between visually assessed and computed colour differences [25]. It is currently the standard figure of merit for this problem [26,27]. $STRESS$ has some interesting properties that encourage us to use it to measure the agreement between the computed image quality and the observers scores. The most relevant one from a theoretical point of view is the possibility of applying statistical significance tests to it. That is, the possibility to figure out through it, and up to a certain degree of confidence, whether the performance of two IQMs can be considered significantly different, from a statistical point of view, or not.

The aim of this paper is two-fold: First, to study the appropriateness of the $STRESS$ index to measure the agreement between the computed and observed image quality scores. Second, to propose a modification of $STRESS$ which incorporates the observers scores' variability or uncertainty in a natural way. A version of $STRESS$, called $WNSTRESS$ (Weighted Normalised STRESS) [25–27], has been previously proposed for this purpose. However, a new modification, called $USTRESS$ (Uncertainty STandardized REsidual Sum of Squares), that uses the standard deviation of the observed scores as a measure of observers' scores variability/uncertainty, is proposed. Incorporating this uncertainty aims to reflect that errors committed by IQMs are more or less important depending on the corresponding observers' scores variability. $USTRESS$ holds almost all of the properties of $STRESS$, including that it is also possible to apply statistical significance tests based on it. In this paper, a comparison of the performance of $STRESS$, $WNSTRESS$, $USTRESS$ and the Pearson, Spearman and Kendall coefficients, using both synthetic datasets and a recent visual image quality evaluation dataset, is carried out. Moreover, we propose a different way to perform statistical significance tests, which better reflects the influence of the uncertainty in the score values.

Fig. 1 shows a flowchart of the whole methodology carried out in this work, for the case of one IQM. Therefore, this process would be repeated for as many IQMs as needed to be compared against. First of all, the observers have analysed each image in the database and assigned opinion scores with a certain average and standard deviation. On the other hand, the IQM for each image in the database is obtained. Then, a measure of agreement is obtained comparing the opinion score with the corresponding IQM. Finally, a ranking of the agreement measures is made to, in turn, rank the IQMs according to their performance.

2. $STRESS$: Standardised residual sum of squares

In multi-dimensional scaling [28,29], loss functions are used to characterise the differences between two vectors (or objects, in general). When these vectors represent groundtruth information (observers' scores) and predicted data (computed scores or IQMs in our context), the closeness between them is interpreted as a measure of approximation quality for the prediction. From now onward, we will only use the terms groundtruth and predicted data, instead of observers' and computed scores.

Thus, in multi-dimensional scaling, the usual loss function is the so-called normalised (or Kruskal's) $STRESS$, which can be defined in different equivalent ways. One of them is the following:

$$STRESS(\mathbf{G},\mathbf{P}) = \left(\frac{\sum_{i=1}^{N} (F_p P_i - G_i)^2}{\sum_{i=1}^{N} G_i^2} \right)^{\frac{1}{2}} \quad (1)$$

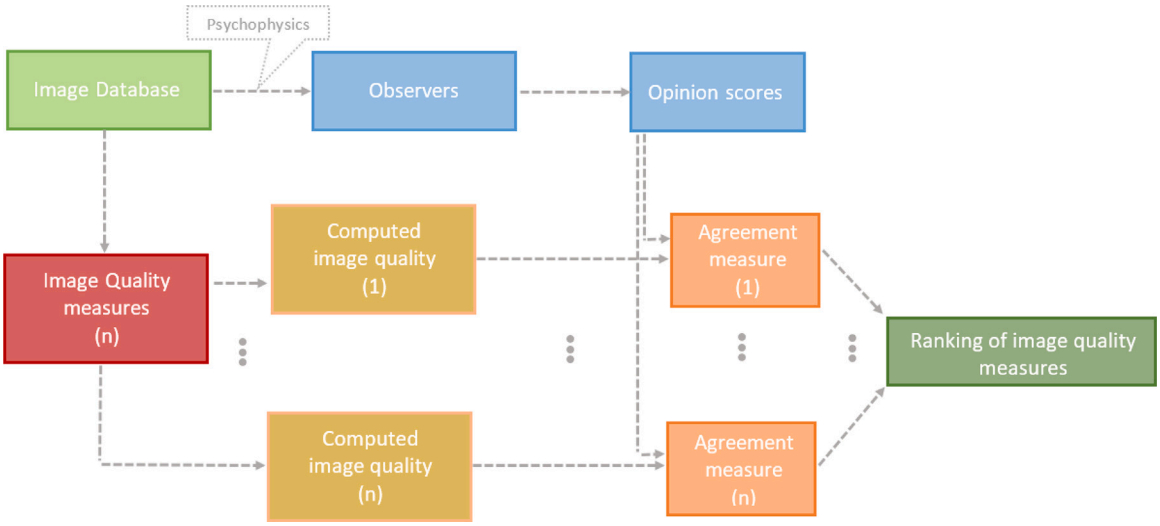

Fig. 1. Flowchart or the methodology proposed in this paper, for one particular agreement measure.

where **G** and **P** are N component vectors denoting groundtruth and predicted data respectively. N here represents the number of image quality scores either determined by visual experimentation or computed with a given IQM. F_p is a non-arbitrary scaling factor determined to minimise the value of the loss function for **P** in relation to **G**. F_p can be analytically determined as:

$$F_p = \frac{\sum_{i=1}^{N} P_i G_i}{\sum_{i=1}^{N} P_i^2} \tag{2}$$

2.1. $WNSTRESS$: Weighted normalised $STRESS$

In a comparison of predicted and groundtruth scores, related to the measurement of the quality of an image, the measures defined up to now do not currently reflect the uncertainty that the dispersion associated to the scores convey. This seems important, because a high dispersion in the scores for an image would indicate that this image might be difficult to assess by the observers, which in turn may bias the results regarding the assessment of the correlation between these quantitative and qualitative measures. For example, a difference equal to m units between computed and groundtruth data should be interpreted differently when the variability of the groundtruth (given in terms of the standard deviation) is low in relation to when it is high.

$STRESS$ can be *adapted* to take into account the contribution of individual comparison pairs through specific weights [25]. Weighted normalised $STRESS$ (or $WNSTRESS$) can be defined as:

$$WNSTRESS(\mathbf{G},\mathbf{P}) = \left(\frac{\sum_{i=1}^{N} \omega_i (\hat{F}_p P_i - G_i)^2}{\sum_{i=1}^{N} \omega_i (G_i)^2} \right)^{\frac{1}{2}} \tag{3}$$

where ω_i is a measurement of the goodness of the data and indicates the weight applied to each *pair*, and $\hat{F}_p$ would have the same values as for $STRESS$, given by Eq. (2). In particular, we would consider: $\omega_i = \frac{1}{\sigma_i^2}$, being σ_i the standard deviation of the groundtruth data of the ith image.

2.2. Statistical significance tests for $STRESS$

It is easy to see that the numerator of Eq. (1) is just a classical Euclidean distance between two vectors, **G** and **P**, one of them appropriately re-scaled (by F_P), i.e. $\sum_{i=1}^{N}(F_P P_i - G_i)^2 = (F_p \cdot \mathbf{P} - \mathbf{G}) \cdot \mathbb{I} \cdot (F_p \cdot \mathbf{P} - \mathbf{G})^T$, where T means transpose, and $\mathbb{I}$ is the $N \times N$ identity matrix in this case. It is not necessary to introduce the identity matrix here, but it is included to easily see the extension of the $STRESS$ measure proposed at the beginning of Section 3.

Therefore, the residual variance of the differences is defined as:

$$V = \frac{\sum_{i=1}^{N}(F_p P_i - G_i)^2}{N-1} \tag{4}$$

which, by the central limit theorem, can be stated to follow a chi-squared distribution with $N-1$ degrees of freedom [28], if N is large enough. It is important to consider that this property would also hold if another semidefinite positive matrix was considered, instead of the identity matrix ($\mathbb{I}$).

Now, for the same groundtruth vector, **G**, given two different prediction vectors $\mathbf{P_1}$ and $\mathbf{P_2}$ their corresponding V_1 and V_2 values can be computed using Eq. (4). Their ratio:

$$F_{test} = \frac{V_1}{V_2},$$

follows, by definition, the distribution of an F variable [29]. It is easy to see that

$$\frac{V_1}{V_2} = \frac{STRESS(\mathbf{G},\mathbf{P}_1)^2}{STRESS(\mathbf{G},\mathbf{P}_2)^2}.$$

As proposed in [25], and using F_{test}, we can formulate the null hypothesis that $\mathbf{P_1}$ and $\mathbf{P_2}$ have no significant differences in predicting **G**. This hypothesis must be rejected when $F_{test} < \frac{1}{F_C}$ or $F_C < F_{test}$, where $\frac{1}{F_C}$ and F_C are the critical bounds of the two-tailed F distribution with a certain (usually 95%) confidence level and $(N-1, N-1)$ degrees of freedom. Notice that F_C is determined as the value that lies in the 97.5% percentile of the F-distribution (and consequently $\frac{1}{F_C}$ is the value that lies in the 2.5% percentile) so that 95% of the samples fall in the interval $\left[\frac{1}{F_C}, F_C\right]$ and 5% out of it.

Consequently, using F_{test}, we may conclude that the predictions $\mathbf{P_1}$ and $\mathbf{P_2}$ are equal ($F_{test} = 1$), insignificantly different ($\frac{1}{F_C} \leq F_{test} \leq F_C$), or significantly different ($F_{test} < \frac{1}{F_C}$ or $F_C < F_{test}$). In the latter case, the one that has the lowest value of V (or $STRESS$) is significantly better than the other.

It is important to note that this statistical significance test cannot be applied for $WNSTRESS$ because it uses the same F_P value that $STRESS$ and consequently the values given by $WNSTRESS$ are not guaranteed to be minimal. In this scenario, statistical significant tests make no sense because they would be biased by F_P.

2.3. A new hypothesis test for $STRESS$

As an alternative way to indicate when two measures are statistically different with a 95% confidence value, we propose to perform a different (albeit related) statistical test. The new test aims to study the null hypothesis (H_0) that $STRESS$ is lower for a given IQM i, than for a given IQM j. Notice that since $STRESS$ should be minimised, this can be interpreted as IQM i being better than IQM j. For this, we compute the $p-$value associated with this hypothesis that represents the probability of error we would incur if we reject the null hypothesis that $STRESS(i) < STRESS(j)$. In particular, the $p-$value is computed as the percentile of the F-distribution where $F_{test} = \frac{V_j}{V_i}$, falls. A very small $p-$value implies that we should reject the null hypothesis, because the risk of being mistaken is also low and, therefore, conclude that IQM j is significantly better than IQM i.

3. $USTRESS$: Uncertainty $STRESS$

We develop in this Section an alternative measure to $STRESS$ and $WNSTRESS$ that is able to consider the uncertainty or variability of the groundtruth data (remember, we assume that it corresponds to the observer's scores) while keeping some of the $STRESS$ properties, particularly the statistical inferences. As explained in detail in the following, we propose to use the Mahalanobis distance instead of the Euclidean distance used in $STRESS$. First, we review the definition of this distance and what it intends to represent.

3.1. The Mahalanobis distance

Let us consider that we have a group of M data points ($\mathbf{X} \equiv \{\mathbf{x}_i\}, i = 1, \dots, M$) in an $N-$dimensional space that are distributed in a natural way, so that the dispersion in each dimension is different and there may exist some correlations among some of the dimensions. A simple way to account for these two facts is to use the variance in each dimension to characterise the first one of them, and the covariance between every dimension pair for the second. All this information is collected in the covariance matrix, $\mathbf{\Sigma}$, associated with the cloud point distribution. Now, to measure the distances between points in the cloud taking into account their natural distribution, the Mahalanobis distance [30–33] proposes to use the inverse of the $\mathbf{\Sigma}$ as the scalar inner product matrix of the $N-$dimensional vector space. By doing this, the distance compensates for the vector differences that are indeed related to the natural distribution of the dataset by assuming that it follows a Gaussian distribution whose covariance matrix is $\mathbf{\Sigma}$. Thus, distances between two points, $\mathbf{x}_i, \mathbf{x}_j$, in the cloud are given by:

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{5}$$

where the superscript T denotes matrix transpose.

It can be seen that since $\mathbf{\Sigma}$ is a (nonsingular) covariance matrix, it is positive definite and hence $d^2(\mathbf{x}_i, \mathbf{x}_j)$ is a metric, in fact [31]. This is the so-called Mahalanobis distance between two data points, and it considers how the variables are dispersed in the space and how they correlate with each other. It is important to emphasise that the use of $\mathbf{\Sigma}^{-1}$ allows for the existence of different scales for the variables, and for nonzero correlations between them. Its quadratic form (Eq. (5)) has the effect of transforming the variables to an uncorrelated standardised variant of them. More in detail, the Mahalanobis distance operates by measuring distances in the context of a multivariate normal distribution characterised by $\mathbf{\Sigma}$ and it is known that the sum of squared residuals of the metric (squared differences) follows a chi-squared distribution [34].

To show how the metric works, it is quite illustrative to define the *locus* of points that are at the same distance from the centroid of the group by the following equation:

$$d^2(\mathbf{x}_i, \mu) = (\mathbf{x}_i - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mu) = K \tag{6}$$

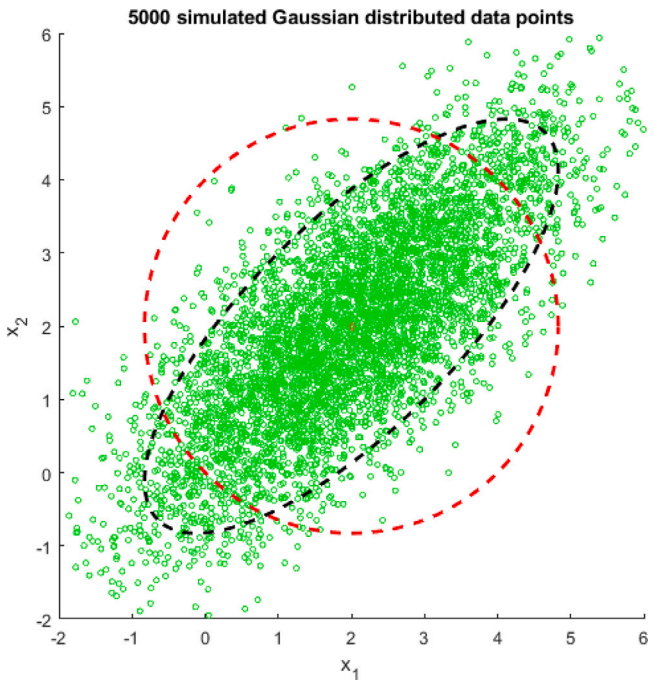

Fig. 2. *Toy* example showing the *benefits* of using the Mahalanobis distance: 5000 2D Gaussian distributed data points (green points) are plotted, whose centroid vector and covariance matrix are: $\mu = (2, 2)$ and $\Sigma = \left(\begin{smallmatrix} 2 & 1.5 \\ 1.5 & 2 \end{smallmatrix}\right)$, respectively. The black dotted curve is the so-called isodensity curve (group of points whose probability density belong to the distribution is the same) for this Gaussian distribution, which corresponds with a set of points at equal Mahalanobis distance from the centroid. This particular locus covers inside the 95% of points of the distribution as the length of each one of the two orthogonal axes corresponds to two times the standard deviation of the distribution in each orthogonal direction. The red dotted curve is the isodensity contour plot for the case of a Gaussian distribution with $\mu = (2, 2)$ and $\Sigma = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$, which corresponds with a set of points at equal Euclidean distance from the centroid.

where μ denotes the centroid and $K > 0$ determines the distance to all points from the centroid.

Fig. 2 shows a toy example of the application of the Mahalanobis distance. Let us consider an $N-$dimensional multivariate Gaussian distribution. Its probability density would be given by:

$$p(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2 \times \pi)^{N/2} \times \sqrt{|\Sigma|}} e^{\left[-\frac{1}{2}[(x-\mu)^T \Sigma^{-1} (x-\mu)]\right]} \tag{7}$$

We can see in Eq. (7) how the Mahalanobis distance appears in the exponent of this probability density function. Fig. 2 shows the plot of a group of 5000 points generated following a 2D Gaussian distribution, with $\mu = (2, 2)$ and $\Sigma = \left(\begin{smallmatrix} 2 & 1.5 \\ 1.5 & 2 \end{smallmatrix}\right)$. Fig. 2 also shows the so-called isodensity curve (group of points whose probability density is the same) for this Gaussian distribution (black dashed line) that corresponds with a set of points at equal Mahalanobis distance from the centroid. The length of the two orthogonal axes of the elliptic curve corresponds to two times the standard deviation of the distribution in those directions which implies that the curve contains the 95% of the cloud. Finally, Fig. 2 also shows (red dashed line), for comparative purposes, the isodensity contour plot for the case of a Gaussian distribution with $\mu = (2, 2)$ and $\Sigma = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$, i.e., a diagonal covariance matrix whose main diagonal elements are equal to those of the previous distribution. The use of the covariance matrix $\Sigma = 2 \cdot \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ instead of $\Sigma = \left(\begin{smallmatrix} 2 & 1.5 \\ 1.5 & 2 \end{smallmatrix}\right)$ in the assessment of the distance between any pair of points in Fig. 2 would be equivalent to the strict application of the Euclidean distance between them (with a factor of 2). We can see how the inclusion of the covariance matrix in the assessment of the probability distribution correctly reflects the nature of the distributed data themselves. Therefore, the Mahalanobis distance (taken into account in Eq. (7)) helps in modelling how data distributes in the feature space.

3.2. $USTRESS$

The main reason of the proposal of the $USTRESS$ index would be to include groundtruth data variability when measuring the agreement with a set of predicted data, IQMs in our case, while keeping some important properties of the $STRESS$ index. For this, we propose to replace the Euclidean distance in $STRESS$ (Eq. (1)) with the Mahalanobis distance, because this distance is able to account for data variability. Therefore, $USTRESS$ can be defined as:

$$USTRESS(\mathbf{G},\mathbf{P}) = \left(\frac{\sum_{i=1}^{N} \left((F_p P_i - G_i)^T \boldsymbol{\Sigma}^{-1} (F_p P_i - G_i) \right)}{\sum_{i=1}^{N} G_i^2} \right)^{\frac{1}{2}} \tag{8}$$

However, this replacement implies, in practical terms, a high dependency on how the matrix $\boldsymbol{\Sigma}$ is defined. For our case, we aim to be able to incorporate the uncertainty associated with the dispersion of the groundtruth data. Therefore, we should keep in mind that:

- Variability is only related to groundtruth information: different observers may give different scores for the same image. However, the score, as computed by a particular IQM is a single value, which obviously has no dispersion.
- The predicted data values are considered independent among them, and therefore the corresponding covariance matrix should be of diagonal nature.

Therefore, we propose to use a particular type of covariance matrix, of diagonal nature, where each element in the so-called *main diagonal* part of the covariance matrix, should reflect the variability of the corresponding groundtruth data (observers' scores), for which we choose the variance of the scores for each stimuli. Thus, we set $\boldsymbol{\Sigma}$ as a diagonal squared matrix, where the ith position in the diagonal matrix, $\boldsymbol{\Sigma}_{ii}$, is equal to the variance of the groundtruth for the stimuli i, σ_i^2. This is a feasible alternative to the use of the sample variance–covariance matrix which is commonly used [30–33]. It is worth to note that how good variance is to characterise data dispersion depends on how many data have been sampled. So, while we can use the same approach for different datasets, results are more meaningful for datasets with more observers' scores, as it is logical.

We must note that the matrix proposed above $\boldsymbol{\Sigma}$ is positive semidefinitive, and therefore the corresponding residual variances are chi-square distributed [31,34]. Therefore, the ratio of variances related to different predicted data (IQMs) also follows, by definition, the distribution of an F variable [29]. Therefore, $USTRESS$ holds this property of $STRESS$ and the statistical tests described in Sections 2.2 and 2.3 can also be used for $USTRESS$.

Finally, using this covariance matrix the expression that defines $USTRESS$ simplifies to:

$$USTRESS(\mathbf{G},\mathbf{P}) = \left(\frac{\sum_{i=1}^{N} \left(\frac{\tilde{F}_p P_i - G_i}{\sigma_i} \right)^2}{\sum_{i=1}^{N} G_i^2} \right)^{\frac{1}{2}} \tag{9}$$

where, following the same mathematical strategy defined in [25], $\tilde{F}_p$ is analytically found to minimise the value of the loss function as:

$$\tilde{F}_p = \frac{\sum_{i=1}^{N} \frac{P_i G_i}{\sigma_i^2}}{\sum_{i=1}^{N} \left(\frac{P_i}{\sigma_i} \right)^2}. \tag{10}$$

Thus, we come up with a natural expression to introduce this uncertainty, where the differences between groundtruth and predicted data are tuned with respect to the corresponding groundtruth variability. In this regard, it is easy to see that the differences associated with the data with $\sigma_i < 1$ will be amplified, whereas when $\sigma_i > 1$ they will be attenuated. In the next section, we will analyse this point deeper and we will see its consequences in practical cases and in relation to the application of statistical significance tests. Note that if $\sigma_i = 1$ for all i, then $USTRESS$ and $STRESS$ are equal.

4. Experimental results

4.1. Synthetic datasets

In this section we perform synthetic experiments that allow us to characterise the properties and performance of $STRESS$, $WNSTRESS$ and $USTRESS$, which we compare to classical Spearman, Pearson, and Kendall's correlation coefficients, and the state of the art measure PWRC [22]. For this, a dataset of groundtruth and predicted data was generated using random values. In this way, we can conveniently control and change the input data to analyse the behaviour of the indexes against these changes in the data. In particular, we have generated 500 pairs of groundtruth and predicted data using a uniformly distributed probability function in the $[1, 5]$ interval. This is typically the application range of these indexes, where values close to 0 are clearly below the perception threshold, and are therefore not interesting.

Initially, this random generation would also provide random results for the above cited correlation measures. From these data, we will study how the agreement measures behave when progressively improve the agreement between groundtruth and prediction data. In addition, we will analyse how the introduction of outliers affects the corresponding correlation indexes. Finally, we will study the influence of different types of uncertainties of the groundtruth data, above these indexes.

First, we started by reducing in an increasing way the initial difference between the ground truth and the predicted data in each pair by modifying the predicted data towards the groundtruth in a fixed percentage of each difference ($|G_i - P_i|, i \in \{1, \dots, N\}$) from 0% to 100% in steps of 10% so that, eventually, we obtain perfect data agreement. No uncertainties of the groundtruth are considered in this stage, thus $WNSTRESS$, $USTRESS$ and $PWRC$ will not be included at the moment. We ran this experiment five times with different random initial values. Fig. 3 shows the average results and standard deviations (the standard deviation is multiplied by 3 for visualisation purposes) between the 5 experiments, provided for $STRESS$, Pearson, Spearman and Kendall coefficients. For clarity of presentation, all measures have been re-scaled in the interval $[0, 100]$ as shown in the legend of Fig. 3. It is clear that the curve for $STRESS$ is almost linear while the ones for Pearson and Spearman are highly nonlinear. In the case of Kendall, linearity is higher than Pearson and Spearman but lower than $STRESS$. In particular, it is especially interesting to note that when differences between predicted and groundtruth data have been reduced 80% or more, the Pearson and Spearman correlations have little sensitivity in this range, while $STRESS$ has the same sensitivity in every reduction step. Having the same sensitivity makes it easier to make the comparison between the values obtained for different IQMs over the same dataset, or alternatively the values obtained by the same IQM over different datasets. The differences between the behaviour of two IQMs would be quite small inside this area around 80% of agreement, thus a high sensitivity of the indexes is very interesting. The behaviour of Kendall in this region is fine, but it is not as linear as $STRESS$. This region of good agreement between groundtruth and predicted data is the usual case in most of the applications of these indexes. As we are not considering the uncertainties at this time, the behaviour of $WNSTRESS$ and $USTRESS$ is the same as $STRESS$.

Secondly, using the 500 pairs of random values for G and the P approached at 80%, which represents realistic values in applications, and also repeating five times the computations, random outliers are introduced into the generated data. In particular, each outlier corresponds to multiplying by a factor of 10 one pair prediction value, chosen randomly. That is, we arbitrarily considered an outlier as a change of one order of magnitude. We proceeded by gradually introducing one by one more outliers, from one to a number of 10, which corresponds from 0.2% to 2% of the whole dataset. Fig. 4 shows the relative worsening, with respect to their initial values without outliers, observed for $STRESS$, Pearson, Spearman and Kendall coefficients (normalised

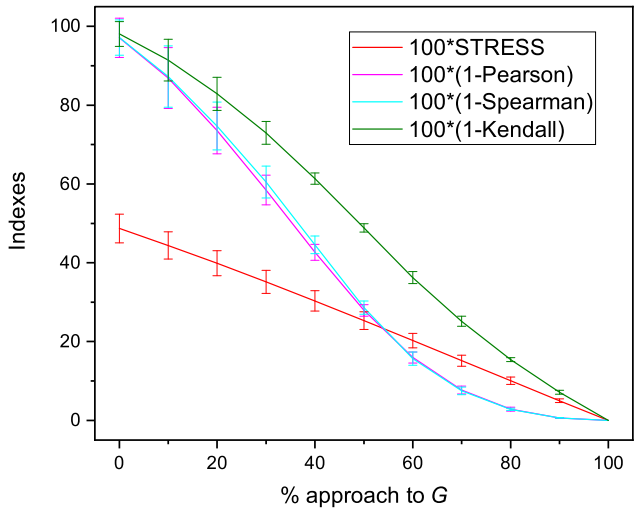

Fig. 3. $STRESS$ values, and Pearson, Spearman and Kendall correlations when increasingly reducing the differences between random groundtruth data and random predictions.

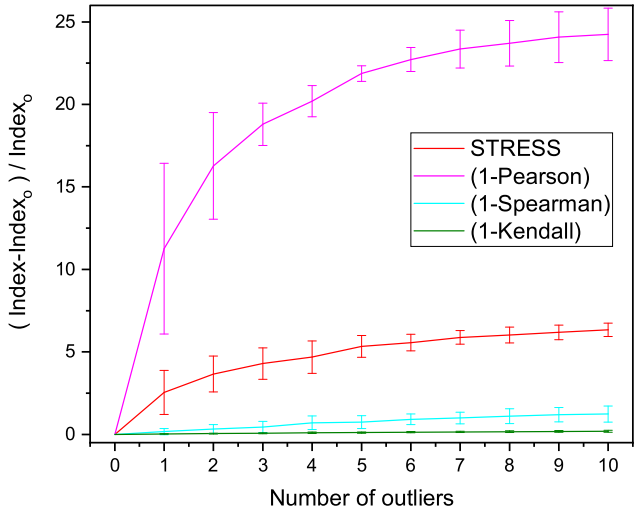

Fig. 4. Relative worsening of $STRESS$ values, and Pearson, Spearman and Kendall correlations when increasing the number of outliers in the data set.

in the range $[0, 100]$ as commented above), when introducing one by one the outliers while keeping the previous ones. Specifically, Fig. 4 shows the average and standard deviations (in this case without any factor) of the relative worsening. We can see that Kendall and Spearman are insensitive to outliers, while Pearson and $STRESS$ are not. The worsening ratio has an up to 25% increase for 10 outliers in the case of Pearson and 5% for $STRESS$. On the one hand, the insensitivity to outliers would be seen as a higher robustness of the index, but makes it impossible to detect outliers. On the other hand, it is expected that a worsening of 2% of the whole dataset must have an effect on the index. Again, with no uncertainties in the groundtruth values, the $WNSTRESS$ and $USTRESS$ values are the same as $STRESS$.

Third and last, we analyse the behaviour of $WNSTRESS$ and $USTRESS$ against $STRESS$ and PWRC and Pearson, Spearman and Kendall coefficients. As we have seen above, $WNSTRESS$ and $USTRESS$ will inherit many of the properties of $STRESS$, as they are based on it. However, when the uncertainties of the groundtruth values are considered, the results of $WNSTRESS$, $USTRESS$ and $STRESS$ are different, because each pair, groundtruth-predicted value, counts

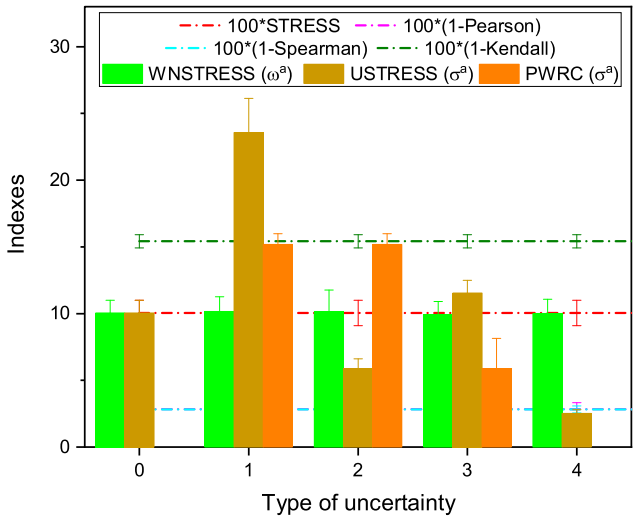

Fig. 5. $STRESS$, $WNSTRESS$, $USTRESS$ and PWRC values, and Pearson, Spearman and Kendall correlations with uncertainties σ^a with four types in the groundtruth data.

differently in the final sum. In usual application of these indexes, uncertainties of the groundtruth correspond to inter- and intra-observers variability in the scores, and can be computed as their standard deviation. The starting point is again 500 pairs of random values for G in the interval $[1, 5]$ and initial P values approached at 80% to G, for the sake of realistic values in applications. This time the value of the uncertainties of the groundtruth values are also considered, through ω_i in Eq. (3) and σ_i in Eqs. (9) and (10), where ω and σ are vectors of 500 values. Note that, if the uncertainties σ_i are considered as the standard deviation between groundtruth data, for consistency $\omega = \frac{1}{\sigma_i^2}$.

To analyse the influence of these uncertainties σ_i has been modelled in two different ways, being ω the inverse of the squared in each case:

- σ^a: σ_i are random numbers using a uniformly distributed probability function centred in different values with different widths: (1) centred in 0.5, width 0.5, i.e. values lower than 1; (2) centred in 2, width 2, i.e. values higher than 1; (3) centred in 1, width 1, i.e. values lower and higher than 1; and (4) centred in 4, width 1, i.e. higher values.
- σ^b: σ_i is set as a percentage, between 10% and 100%, of the groundtruth value.

In both cases, σ^a and σ^b, a first step is included corresponding to values of σ equal to 1, which represents the case when WNSTRESS and USTRESS coincide in value with STRESS just to show how much they change when different uncertainties are introduced.

The two considered σ include a progressive uncertainty of the data to study their effects on the indexes. For σ^a the 4 cases corresponds respectively to uncertainties below 1, above 1, above and below 1 and above 1 but greater.

Fig. 5 shows the results for uncertainties σ^a and Fig. 6 shows the results for σ^b. Once again it is plotted the mean of five different experiments, corresponding the error bars to the standard deviation between the 5 (multiplied by 3 for visualisation purposes). Obviously, $STRESS$, Pearson, Spearman and Kendall correlations are insensitive to the uncertainty in the data, and have the same value in all the cases. Just this is the reason why $WNSTRESS$ and $USTRESS$ have been introduced. In both figures the plotted values for Pearson (2.85 ± 0.47) and Spearman ($2.81 \pm 0{,}26$) are almost identical and overlap in the graphs. In the case of Kendall the plotted value is (15.42 ± 0.50) and in the case of $STRESS$, its value is 10.05 ± 0.95 in all the cases.

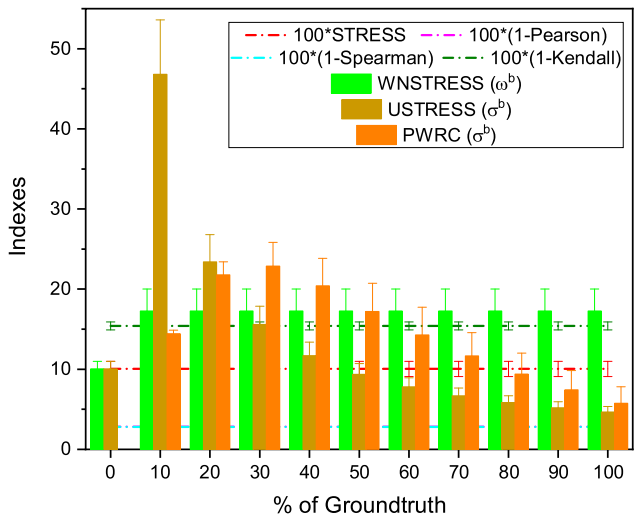

Fig. 6. $STRESS$, $WNSTRESS$, $USTRESS$ and PWRC values, and Pearson, Spearman and Kendall correlations with uncertainties σ^b with 10 different degrees in the groundtruth data.

Table 1
Results of the statistical significant test for $STRESS$ under the null hypothesis (H_0) that the $STRESS$ value for IQM in row i ($STRESS(i)$) is insignificantly different than the $STRESS$ value for IQM in column j ($STRESS(j)$), for the CID:IQ dataset.

STRESS	iCAM	FCSS	SSIM	MSSSIM	CMSSIM	FSIMc	MSE	RMSE	PSNR	NCD
iCAM										
FCSS										
SSIM										
MSSSIM										
CMSSIM										
FSIMc										
MSE										
RMSE										
PSNR										
NCD										

Table 2
Results of the statistical significant test for $USTRESS$ under the null hypothesis (H_0) that the $USTRESS$ value for IQM in row i ($STRESS(i)$) is insignificantly different than the $USTRESS$ value for IQM in column j ($USTRESS(j)$), for the CID:IQ dataset.

USTRESS	iCAM	FCSS	SSIM	MSSSIM	CMSSIM	FSIMc	MSE	RMSE	PSNR	NCD
iCAM										
FCSS										
SSIM										
MSSSIM										
CMSSIM										
FSIMc										
MSE										
RMSE										
PSNR										
NCD										

Initially, with $\sigma = 1$, $WNSTRESS$ and $USTRESS$ give the value of $STRESS$. Note that PWRC cannot be computed for this case. It should be noted that $WNSTRESS$ is not affected by the uncertainty of the data for each case of ω. In the case of ω^a (Fig. 5) the values are almost equal than $STRESS$, being slightly higher in the case of ω^b (Fig. 6), but also insensitive to changes.

Thus, only $USTRESS$ and PWRC are sensitive to uncertainties in the groundtruth data. In the case where all uncertainties are below 1 (Fig. 5, type 1) the value of $USTRESS$ clearly increases and, conversely, decreases in the cases of all uncertainties above 1 (Fig. 5, types 2 and 4). With the uncertainties above and below 1 of type 3 we obtain intermediate values. In the case of the PWRC index, it can be computed neither with the uncertainties type 0 or 4. All uncertainties below 1 of above 1 give the same value for PWRC. Only uncertainties below and above 1 make the index decrease. In conclusion, $USTRESS$ is much more sensitive to these different types of uncertainties.

By modelling the uncertainties in the groundtruth data by σ^b, a similar behaviour can be seen in Fig. 6. In the comparison of $USTRESS$ with $STRESS$ (the value of $USTRESS$ with all $\sigma = 1$), small uncertainties (Fig. 6, 10% and 20%) worsen the final agreement between groundtruth and predicted data, while large uncertainties (Fig. 6, above 50%) make that $USTRESS$ better agreement is determined. The behaviour of $USTRESS$ is quite linear with the progressive change in σ^b. In this case, $WNSTRESS$ has an increase from 10 to 17, which means that it is sensitive to the introduction of weights but not to their change. For these type of uncertainties PWRC is sensitive, with a behaviour similar to $USTRESS$ for large uncertainties (i.e. about 10%), but different for small uncertainties. Anyway the changes in the value of PWRC are smaller than for $STRESS$.

Summarising, $USTRESS$ is sensitive to any type of uncertainty in the groundtruth data, independently of the type and degree. In addition, it is always possible to compute $USTRESS$, while the assessment of PWRC has some restrictions.

4.2. Image quality scores dataset

In this section, the performance of Spearman, Pearson, and Kendall's correlation coefficients, and $STRESS$, $WSTRESS$, $USTRESS$ and PWRC [22] measures are compared. All of them are used to analyse the agreement between values predicted by IQMs and groundtruth scores for two real experimental datasets: the Colourlab Image Database: Image Quality (CID:IQ) [35] and the KONIQ-10K (KONIQ) [36] database. For CID:IQ, we will use full-reference IQMs, while no-reference IQMs will be used for KONIQ.

4.2.1. CID:IQ

The CID:IQ dataset contains 23 pictorial images selected as the reference images with 6 different distortions, over 5 levels. The applied distortions are as follows: JPEG compression, JPEG2000 compression, Poisson noise, blurring, and two gamut mapping algorithms. These images were evaluated by a group of 17 observers, whose image quality scores were used as groundtruth data. It is required to have a dataset where the variances of subjective scores are given, which results in datasets such as LIVE [37], TID2008 [38], TID2013 [39] cannot be used as they do not report the variance. To predict the image quality data values, the 10 state-of-the-art full-reference IQMs listed in Section 1 have been considered.

In Fig. 7, we compare the agreement between the predicted data (IQMs) and the average groundtruth data (observers scores), given by Spearman, Pearson, and Kendall's correlation coefficients, and $STRESS$, $WSTRESS$, $USTRESS$ and PWRC [22] measures for the CID:IQ dataset. Standard deviation of the groundtruth data (i.e., observers' scores) have been used in $WSTRESS$ and $USTRESS$ as explained in Section 4.1. We can see that there are some ranking differences between the 5 indexes but all agree on identifying SSIM, FCSS, FSIMc, and PSNR as those showing the best performance. However, the question that arises is whether the differences among these measures are significant or not. We will try to shed some light on this by using the statistical significance tests described above, which are available only for $STRESS$ and $USTRESS$. As we can see in Fig. 7, although the rankings of IQMs do agree for $STRESS$, $WSTRESS$ and $USTRESS$, their absolute values differ, and the statistical significance tests show different results for $STRESS$ and $USTRESS$ as we will see below.

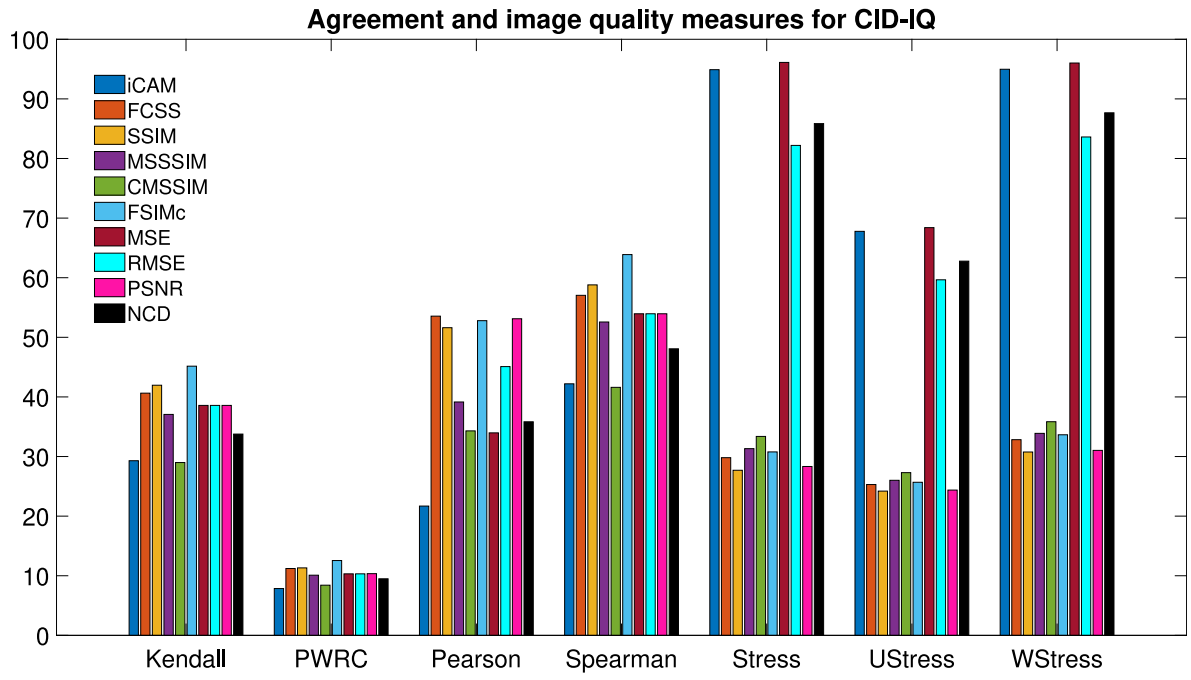

Fig. 7. Agreement measures computed by each one of the indexes in the comparison list for the CID:IQ dataset. Higher values for Kendall, PWRC, Pearson and Spearman indicate better performance, while for $STRESS$, $USTRESS$ and $WSTRESS$, lower values indicate a better performance.

Table 3
p–value for the null hypothesis (H_0) that the $STRESS$ value for IQM in row i ($STRESS(i)$) is lower than the $STRESS$ value for IQM in column j ($STRESS(j)$), for the CID:IQ dataset.

STRESS	iCAM	FCSS	SSIM	MSSSIM	CMSSIM	FSIMc	MSE	RMSE	PSNR	NCD
iCAM	0.500	0.000	0.000	0.000	0.000	0.000	0.631	0.000	0.000	0.004
FCSS	1.000	0.500	0.027	0.903	0.998	0.799	1.000	1.000	0.091	1.000
SSIM	1.000	0.973	0.500	0.999	1.000	0.997	1.000	1.000	0.724	1.000
MSSSIM	1.000	0.097	0.001	0.500	0.952	0.322	1.000	1.000	0.004	1.000
CMSSIM	1.000	0.002	0.000	0.048	0.500	0.017	1.000	1.000	0.000	1.000
FSIMc	1.000	0.201	0.003	0.678	0.983	0.500	1.000	1.000	0.015	1.000
MSE	0.369	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.000	0.002
RMSE	1.000	0.000	0.000	0.000	0.000	0.000	1.000	0.500	0.000	0.873
PSNR	1.000	0.909	0.276	0.996	1.000	0.985	1.000	1.000	0.500	1.000
NCD	0.996	0.000	0.000	0.000	0.000	0.000	0.998	0.127	0.000	0.500

Table 4
p–value under the null hypothesis (H_0) that the $USTRESS$ value for IQM in row i ($USTRESS(i)$) is lower than the $USTRESS$ value for IQM in row j ($USTRESS(j)$), for the CID:IQ dataset.

USTRESS	iCAM	FCSS	SSIM	MSSSIM	CMSSIM	FSIMc	MSE	RMSE	PSNR	NCD
iCAM	0.500	0.000	0.000	0.000	0.000	0.000	0.595	0.000	0.000	0.022
FCSS	1.000	0.500	0.121	0.767	0.977	0.652	1.000	1.000	0.163	1.000
SSIM	1.000	0.879	0.500	0.971	0.999	0.940	1.000	1.000	0.574	1.000
MSSSIM	1.000	0.233	0.029	0.500	0.897	0.368	1.000	1.000	0.044	1.000
CMSSIM	1.000	0.023	0.001	0.103	0.500	0.055	1.000	1.000	0.001	1.000
FSIMc	1.000	0.348	0.060	0.632	0.945	0.500	1.000	1.000	0.085	1.000
MSE	0.405	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.000	0.012
RMSE	1.000	0.000	0.000	0.000	0.000	0.000	1.000	0.500	0.000	0.910
PSNR	1.000	0.837	0.426	0.956	0.999	0.915	1.000	1.000	0.500	1.000
NCD	0.978	0.000	0.000	0.000	0.000	0.000	0.988	0.090	0.000	0.500

Subsequently, we run statistical significance tests for $STRESS$ and $USTRESS$, as explained in Section 2.2. The results for the first tests are shown in the Tables 1 and 2. In this case the null hypothesis (H_0) for table position (i, j) is that the $STRESS$ (or the $USTRESS$) value for IQM i is insignificantly different than the $STRESS$ value for IQM j. If this hypothesis can be rejected at a 95% confidence level, then the colour assigned in the table position (i, j) is green, and red otherwise. In particular, for $STRESS$, IQMs SSIM, FCSS, and PSNR do not show statistically significant differences at a 95% confidence level. However, the difference between SSIM and FSIMc is indeed statistically significant. For $USTRESS$, not only are the differences between SSIM, FCSS, and PSNR insignificant but also the differences between those and FSIMc and MSSSIM. In general, we can see that there are more insignificant differences found for $USTRESS$ than for $STRESS$. This happens because in this dataset, overall, standard deviations of observers scores are higher than 1. This, reduces the values in numerator of Eq. (9) and consequently the quotient computed in the F-test gets closer to 1 and more likely to be within the given confidence limits.

However, these results only reflect those cases when the computed ratios for the test move in or out of the critical bounds established for the given confidence interval. We think it is also interesting to look at the variations in the ratios in general. Therefore, we also run the alternative significance test described in Section 2.3. Tables 3 and 4 show in position (i, j) the $p-$value of the contrast hypothesis when comparing the $STRESS$ and the $USTRESS$ values for a pair of IQMs, (i, j), for the CID:IQ dataset. This $p-$value quantifies the risk to be mistaken that we should assume if we reject the null hypothesis. In our case, the null hypothesis is that $STRESS(i) < STRESS(j)$ (or $USTRESS(i) < USTRESS(j)$). Therefore, a very small $p-$value implies that we should reject the null hypothesis, because the risk to be mistaken is as low as the corresponding $p-$value. In our particular case, since a lower $STRESS$ (or $USTRESS$) value means better quality, rejecting the hypothesis $STRESS(i) < STRESS(j)$ implies rejecting the hypothesis that IQM i is better than IQM j. That is, it means exactly the opposite: IQM j is significantly better than IQM i. However, a large $p-$value implies the opposite: the null hypothesis should not be rejected and then IQM i is significantly better than IQM j.

Looking at this table in more detail, we can see that the $p-$values found for $USTRESS$ are less extreme than those for $STRESS$, that is, they are closer to 0.5. Notice that $p-$values close to 0 imply strong evidence that the null hypothesis should be rejected. On the other hand, $p-$values close to 1 imply the opposite, that is, rejecting the null hypothesis implies a high error probability. Then, in our case, when the $p-$values are more extreme it is easier to conclude if a given IQM is significantly better than another. However, when the $p-$values are closer to 0.5 in general, there is less evidence indicating that one metric is significantly better than another, as 0.5 is the value that does not represent any evidence in favour or against the null hypothesis. That is, according to $USTRESS$, the differences are in general less significant, as a consequence of including uncertainty in the groundtruth data. This happens because in this dataset, overall, standard deviations of observers scores are higher than 1, as said before. As an example, in a comparison of the two best performing IQMs, SSIM and PSNR, we can see that they do not have significant performance differences in the first statistical test. In the second, we can see that the probability of error we incur when rejecting the hypothesis that SSIM is significantly better than PSNR is 0.724 for $STRESS$ and 0.574 for $USTRESS$. So, according to $STRESS$ there is slight evidence that SSIM is better than PSNR, but according to $USTRESS$ this evidence is practically negligible as its $p-$value is very close to 0.5.

Last, it is important to note that whatever influence there is in including data uncertainty in the statistical significance tests, this is independent on how well IQMs agree with average observers' scores as the uncertainty of data is commonly independent from its average (except when using some specific probability distributions).

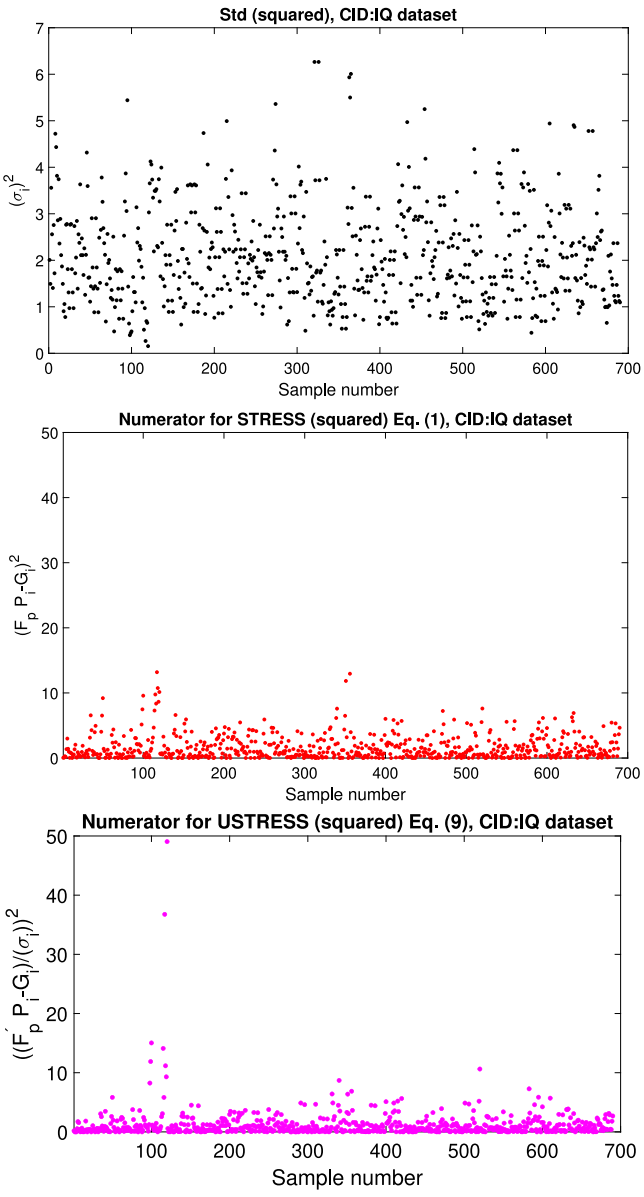

Fig. 8. Comparison of squared sample errors in $STRESS$ and $USTRESS$ in relation to their corresponding scores standard deviations. On the top we show standard deviation of observers scores. The middle and bottom plots are the squared error in numerator of Eqs. (1) and (9), respectively.

We can analyse a bit deeper the behaviour discussed. We can see that the average standard deviations of the groundtruth data in the dataset is about 1.40 and most of them are higher than 1. This means that, according to Eq. (9), in most cases the differences between predicted and groundtruth data are attenuated, which makes the IQMs performace being closer together, and so there are less significant differences among them. This effect is shown in Fig. 8 where we plot for each image in the dataset the standard deviation of the observers' scores and the individual errors found in $STRESS$ and $USTRESS$ that are summarised in Eqs. (1) and (9), respectively. It can be seen that in the case of $USTRESS$, most of the errors are attenuated with respect to $STRESS$. For instance, those for images 95, 321, 363, 365, 454 which are associated to large standard deviations. This means that $USTRESS$ models that errors committed related to cases where observers' scores differ greatly are less important. It also happens that for some images the standard deviation of the scores is very low; 117, 120, 520 and 583 for instance. In these cases, the errors are

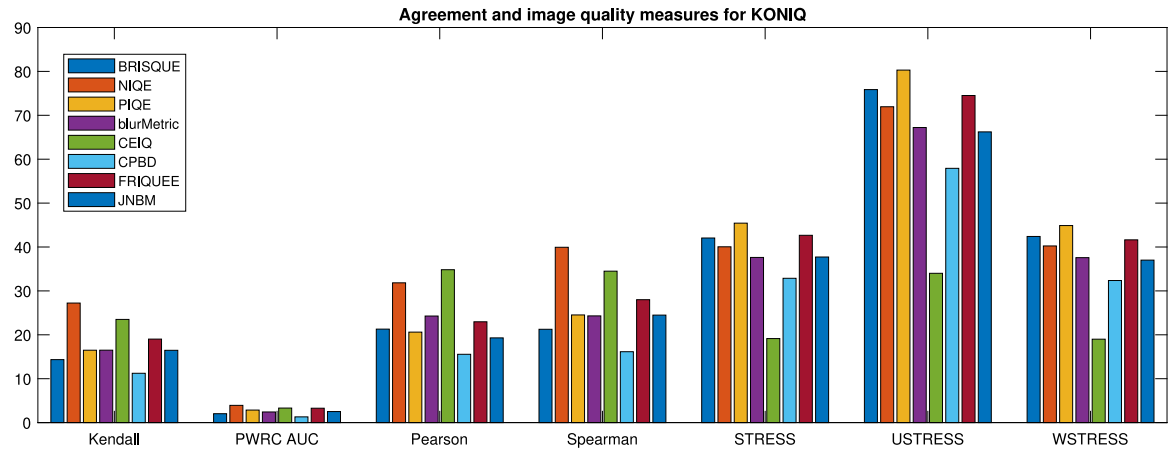

Fig. 9. Agreement measures computed by each of the indexes in the comparison list for the KONIQ dataset. Higher values for Kendall, PWRC, Pearson and Spearman indicate better performance, while for $STRESS$, $USTRESS$ and $WSTRESS$ lower values indicate better performance.

amplified in $USTRESS$ which means that errors committed by IQMs for images where there is good agreement in observers' opinions are more important. We strongly believe that this reasoning makes total sense in general when measuring the agreement between groundtruth and predicted data in the psychophysics field, which is one of the main advantages of $USTRESS$.

4.2.2. KONIQ

The KONIQ dataset consists of 10,073 quality scored images from an online experiment. It has a total of 1.2 million ratings from 1459 observers. KONIQ contains the following distortions: noise, JPEG artifacts, aliasing, lens and motion blur, over-sharpening, wrong exposure, colour fringing, and over-saturation. The experiment was carried out using the standard 5-point absolute category rating scale. We use the following no-reference IQMs for the analysis as mentioned in the introduction: BRISQUE [10], NIQE [11], PIQE [12], blurMetric [13], CEIQ [14], CPBD [15], FRIQUEE [16], and JNBM [17].

The results for all the agreement measures considered can be found in Fig. 9. From this figure we can see that Kendall, Pearson, Spearman and PWRC coincide on pointing NIQE and CEIQ as the best two performing IQMs followed by FRIQUEE, PIQE and blurMetric with slight different orderings. However, STRESS, USTRESS and WSTRESS find CEIQ and CPBD the two best ones, followed by blurMetric and JNBM in the case of STRESS and JNBM and blurMetric in the case of USTRESS and WSTRESS. The fact that JNBM gets better than blurMetric when the score uncertainty is taken into account means that JNBM makes less error than blurMetric in images where the observers agree more (low standard deviation of scores) and/or that blurMetric makes less error than JNBM in images where the observers do not agree that much (high standard deviation). Both situations would be in favour of JNBM when the standard deviation is considered.

Statistical significance tests for STRESS and USTRESS can be found in Tables 5 and 6. In this case, there is no difference between the tests at 95% of confidence level but we do see clear changes in the p-values computed in Tables 7 and 8. For instance, in the comparison between BRISQUE and FRIQUEE which is opposite for STRESS and UTRESS, and also when comparing JNBM and blurMetric: P-value in position (JNBM, blurMetric) is 0.398 for STRESS and 0.932 for USTRESS. So, for STRESS there is a slight evidence of blurMetric being better than JNBM whereas for USTRESS there is a quite strong evidence of the opposite.

Table 5
Results of statistical significant test for $STRESS$ under the null hypothesis (H_0) that the $STRESS$ value for IQM in row i ($STRESS(i)$) is insignificantly different than the $STRESS$ value for IQM in column j ($STRESS(j)$), for the KONIQ dataset.

STRESS	BRISQUE	NIQE	PIQE	blurMetric	CEIQ	CPBD	FRIQUEE	JNBM
BRISQUE								
NIQE								
PIQE								
blurMetric								
CEIQ								
CPBD								
FRIQUEE								
JNBM								

Table 6
Results of statistical significant test for $USTRESS$ under the null hypothesis (H_0) that the $USTRESS$ value for IQM in row i ($STRESS(i)$) is insignificantly different than the $USTRESS$ value for IQM in column j ($USTRESS(j)$), for the KONIQ dataset.

USTRESS	BRISQUE	NIQE	PIQE	blurMetric	CEIQ	CPBD	FRIQUEE	JNBM
BRISQUE								
NIQE								
PIQE								
blurMetric								
CEIQ								
CPBD								
FRIQUEE								
JNBM								

5. Conclusions

We have proposed not only the $STRESS$ measure but an extension of it, $USTRESS$, to assess the performance of image quality metrics. $USTRESS$ incorporates information about the variability of observer's scores. This information is not incorporated in commonly used performance measures, such as Pearson and Spearman correlation coefficients.

Evaluations on synthetic and real datasets have shown that the proposed $USTRESS$ measure has advantageous properties. On the one hand, it is more linear, sensitive, robust and stable than other measures. On the other hand, it allows to apply statistical significance tests to the results found being not only able to rank a series of image quality metrics but to characterise up to what extent the ranking differences are

Table 7
p–value for the null hypothesis (H_0) that the $STRESS$ value for IQM in row i ($STRESS(i)$) is lower than the $STRESS$ value for IQM in column j ($STRESS(j)$), for the KONIQ dataset.

STRESS	BRISQUE	NIQE	PIQE	blurMetric	CEIQ	CPBD	FRIQUEE	JNBM
BRISQUE	0.500	0.000	1.000	0.000	0.000	0.000	0.928	0.000
NIQE	1.000	0.500	1.000	0.000	0.000	0.000	1.000	0.000
PIQE	0.000	0.000	0.500	0.000	0.000	0.000	0.000	0.000
blurMetric	1.000	1.000	1.000	0.500	0.000	0.000	1.000	0.602
CEIQ	1.000	1.000	1.000	1.000	0.500	1.000	1.000	1.000
CPBD	1.000	1.000	1.000	1.000	0.000	0.500	1.000	1.000
FRIQUEE	0.072	0.000	1.000	0.000	0.000	0.000	0.500	0.000
JNBM	1.000	1.000	1.000	0.398	0.000	0.000	1.000	0.500

Table 8
p–value under the null hypothesis (H_0) that the $USTRESS$ value for IQM in row i ($USTRESS(i)$) is lower than the $USTRESS$ value for IQM in row j ($USTRESS(j)$), for the KONIQ dataset.

USTRESS	BRISQUE	NIQE	PIQE	blurMetric	CEIQ	CPBD	FRIQUEE	JNBM
BRISQUE	0.500	0.000	1.000	0.000	0.000	0.000	0.036	0.000
NIQE	1.000	0.500	1.000	0.000	0.000	0.000	1.000	0.000
PIQE	0.000	0.000	0.500	0.000	0.000	0.000	0.000	0.000
blurMetric	1.000	1.000	1.000	0.500	0.000	0.000	1.000	0.068
CEIQ	1.000	1.000	1.000	1.000	0.500	1.000	1.000	1.000
CPBD	1.000	1.000	1.000	1.000	0.000	0.500	1.000	1.000
FRIQUEE	0.964	0.000	1.000	0.000	0.000	0.000	0.500	0.000
JNBM	1.000	1.000	1.000	0.932	0.000	0.000	1.000	0.500

really significant. $USTRESS$ therefore, results in a beneficial measure for the assessment of the performance of image quality metrics. The proposed measure is seen as a complement to existing performance measures.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

S. Morillas and R. Huertas acknowledge the support of Generalitat Valenciana under grant AICO-2020-136. R. Huertas acknowledges the support under the research project FIS2017-89258-P (“Ministerio de Economía, Industria y Competitividad”, “Agencia Estatal de Investigacion”, Spain) along with the European Union FEDER (European Regional Development Funds) support. M. Pedersen acknowledges the support of the Research Council of Norway through the project “Quality and Content: understanding the influence of content on subjective and objective image quality assessment” (project number 324663).

References

[1] M. Pedersen, J.Y. Hardeberg, Full-reference image quality metrics: Classification and evaluation, Found. Trends Comput. Graph. Vis. 7 (1) (2012) 1–80.
[2] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process., 13 (4) 600–612.
[3] https://www.scopus.com/authid/detail.uri?authorId=56984291600.
[4] M.D. Fairchild, G.M. and Johnson, iCAM framework for image appearance, differences, and quality, J. Electron. Imaging, 13 (1) 126–138.
[5] S. Grečova, S. Morillas, Perceptual similarity between color images using fuzzy metrics, J. Vis. Commun. Image Represent. 34 (2016) 230–235.
[6] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multi-scale structural similarity for image quality assessment, in: Invited Paper, IEEE Asilomar Conference on Signals, Systems and Computers, 2003.
[7] M. Hassan, C. Bhagvati, Structural similarity measure for color images, Int. J. Comput. Appl. 43 (14) (2012) 7–12.
[8] Lin Zhang, Lei Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386, Source code is available at http://www4.comp.polyu.edu.hk/cslzhang/IQA/FSIM/FSIM.htm.
[9] K.N. Plataniotis, A.N. Venetsanopoulos, Color Image Processing and Applications, Springer-Verlag, 2000, p. 355, pp 1-45, 51-100, 109-157.
[10] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.
[11] A. Mittal, R. Soundararajan, A.C. Bovik, Making a completely blind image quality analyzer, IEEE Signal Process. Lett. 22 (3) (2013) 209–212.
[12] N. Venkatanath, D. Praneeth, Bh.M. Chandrasekhar, S.S. Channappayya, S.S. Medasani, Blind image quality evaluation using perception based features, in: Proceedings of the 21st National Conference on Communications, NCC, IEEE, Piscataway, NJ, 2015.
[13] Frederique Crete, Thierry Dolmiere, Patricia Ladret, Marina Nicolas, The blur effect: perception and estimation with a new no-reference perceptual blur metric, in: Human Vision and Electronic Imaging XII, 6492, SPIE, 2007, pp. 196–206.
[14] Yuming Fang, Kede Ma, Zhou Wang, Weisi Lin, Zhijun Fang, Guangtao Zhai, No-reference quality assessment of contrast-distorted images based on natural scene statistics, IEEE Signal Process. Lett. 22 (7) (2014) 838–842.
[15] Niranjan D. Narvekar, Lina J. Karam, A no-reference image blur metric based on the cumulative probability of blur detection (CPBD), IEEE Trans. Image Process. 20 (9) (2011) 2678–2683.
[16] Deepti Ghadiyaram, Alan C. Bovik, Perceptual quality prediction on authentically distorted images using a bag of features approach, J. Vis. 17 (1) (2017) 32.
[17] Rony Ferzli, Lina J. Karam, A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB), IEEE Trans. Image Process. 18 (4) (2009) 717–728.
[18] J.L. Rodgers, W.A. Nicewander, Thirteen ways to look at the correlation coefficient, Am. Stat. 42 (1988) 59–66.
[19] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72–101.

[20] M. Kendall, A new measure of rank correlation, Biometrika 30 (1–2) (1938) 81–89, http://dx.doi.org/10.1093/biomet/30.1-2.81.JSTOR2332226.
[21] Objective Perceptual Assessment of Video Quality: Full Reference Television, Video Quality Experts Group, 2004.
[22] Q. Wu, H. Li, F. Meng, K.N. Ngan, A perceptually weighted rank correlation indicator for objective image quality assessment, IEEE Trans. Image Process. 27 (5) (2018) 2499–2513, http://dx.doi.org/10.1109/TIP.2018.2799331.
[23] J.B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Pychometrika 29 (1964) 1–27.
[24] A.P.M. Coxon, The User's Guide to Multidimensional Scaling, Heinemann, 1982.
[25] P.A. Garcia, R. Huertas, M. Melgosa, G. Cui, Measurement of the relationship between perceived and computed color differences, JOSA - A 24 (7) (2007) 1823–1829.
[26] C. Li, Z. Li, Z. Wang, Y. Xu, M.R. Luo, G. Cui, M. Melgosa, M.H. Brill, M. Pointer, Comprehensive color solutions: CAM16, CAT16, and CAM16-UCS, Color Res. Appl. 42 (6) (2017) 703–718.
[27] M. Safdar, G. Cui, Y.J. Kim, M.R. Luo, Perceptually uniform color space for image signals including high dynamic range and wide gamut, Opt. Express 25 (2017) 15131–15151.
[28] J.F. Seely, D. Birkes, Y. Lee, Characterizing sums of squares by their distributions, Amer. Statist. 51 (1) (1997) 55–58, http://dx.doi.org/10.2307/2684696, JSTOR www.jstor.org/stable/2684696.
[29] R.A. Fisher, On the mathematical foundations of theoretical statistics, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 222 (1922) 309–368.
[30] P.C. Mahalanobis, On the generalised distance in statistics, Proc. Nat. Inst. Sci. India 2 (1936) 49–55.
[31] G.J. Mclachlan, Mahalanobis distance, Resonance (1999).
[32] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Tutorial: The mahalanobis distance, Chemometr. Intell. Lab. Syst. 50 (2000) 1–18.
[33] R.G. Brereton, The mahalanobis distance and its relationship to principal components scores, J. Chemometr. 29 (2015) 143–145.
[34] R.G. Brereton, The chi squared and multinormal distributions, J. Chemometr. 29 (2015) 9–12.
[35] X. Liu, M. Pedersen, J.Y. Hardeberg, CID: IQ–a new image quality database, in: International Conference on Image and Signal Processing, 2014, pp. 193–202.
[36] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, Dietmar Saupe, Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment, IEEE Trans. Image Process. 29 (2020) 4041–4056.
[37] H.R. Sheikh, Z. Wang, L. Cormack, A.C. Bovik, LIVE Image Quality Assessment Database Release 2. http://live.ece.utexas.edu/research/quality.
[38] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, TID2008-a database for evaluation of full-reference visual quality assessment metrics, Adv. Mod. Radioelectron. 10 (4) (2009) 30–45.
[39] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.C.J. Kuo, Image database TID2013: Peculiarities, results and perspectives, Signal Process., Image Commun. 30 (2015) 57–77.