27

Using STRESS to compute the agreement between computed image quality measures and observer scores: advantanges and open issues

S. Morillas [♭],[1] P. Latorre-Carmona [♮], R. Huertas[♢], and M. Pedersen[♡]

(♭) Intituto de Matemática Pura y Aplicada, Universitat Politècnica de València.
(♮) Departamento de Ingeniería Informática, Universidad de Burgos.
(♢) Departamento de Óptica, Universidad de Granada.
(♡) Department of Computer Science, Norwegian Unviersity of Science and Technology.

Abstract

Visual quality of color images is studied through costly psychophysical experiments, which are used to record observers quality scores. Visual image quality metrics pursue to maximize the agreement between computed quality and observers scores. Therefore, it is of critical importance to have appropriate measures for this agreement, both for the development and use of the image quality metrics. The most used one is the well known Pearson correlation coefficient while Spearman rank correlation coefficient is also customary used. In this work we explore the use of an alternative metric: The standardized residual sum of squares (STRESS). STRESS has some interesting properties that encourage us to use it for measuring the agreement between computed image quality and observers scores, being the most important one the possibility to run statistical significance tests between metrics. We will compare the performance of STRESS with Pearson and Spearman coefficients using both synthetic datasets as well as a recent visual image quality evaluation dataset. As it will be shown, the performance is different and we found several points in favor of using STRESS along with some interesting open issues.

1 Introduction

Image quality assessment is currently an active research topic due to the ubiquitous use of color images for domestic, scientific and industrial applications. A small proof of this is that the most popular image quality metric developed so far, the structural similarity index (SSIM) [1], nowadays accounts for more than 23K citations according to Scopus [2].
In particular, visual quality of color images is studied through costly psychophysical experiments used to record observers quality scores. Visual image quality metrics aim at maximizing the agreement between computed quality and observers scores. Therefore, it is of critical importance to have appropriate measures for this agreement. The most used one is the well known Pearson correlation coefficient [3] while Spearman rank correlation coefficient [4] is also customary used. In this work we explore the use of an alternative metric: Standardized Residual Sum of Squares (STRESS). This metric was originally employed in multidimensional scaling (MDS) techniques

[1] smorillas@mat.upv.es

[5,6], and later have been extensively used to measure the agreement between visually assessed and computed color differences [7], being the standard figure of merit for this problem [8,9]. STRESS has some interesting properties that encourage us to use it for measuring the agreement between computed image quality and observers scores. The most relevant one from a theoretical point of view is the possibility to apply statistical significance tests. That is, the possibility to figure out, up to a certain degree of confidence, if the performance of two metrics can be considered different enough, from a statistical point of view. In this paper a comparison of the performance of STRESS and Pearson and Spearman coefficients is shown, using both synthetic datasets as well as a recent visual image quality evaluation dataset [11]. As it will be shown, the performance is quite different and we found several practical points in favor of using STRESS along with some interesting open issues.

2 STRESS: Standardized Residual Sum of Squares

In multi-dimensional scaling [12,13], loss functions are used to characterize the differences between two vectors (or objects, in general). When these vectors represent groundtruth and predicted data, the closeness between them is interpreted as a measure of approximation quality for the prediction. The usual loss function is the so-called normalized (or Kruskal's) STRESS, which can be defined in different equivalent ways, one of them the following:

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

where $\mathbf{G}$ and $\mathbf{P}$ are N component vectors denoting groundtruth and predicted data, respectively, and F_P is a non-arbitrary scaling factor determined to minimize the value of the loss function for $\mathbf{P}$ in relation to $\mathbf{G}$. F_P can be analytically determined to be:

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

2.1 Statistical significance tests for STRESS

By looking at the numerator of Eq. 1, we can see that we are just using a classical euclidean distance between two vectors, $\mathbf{G}$ and $\mathbf{P}$, one of them appropriately re-scaled (by F_P). In particular,

$$V = \frac{\sum_{i=1}^{N} (G_i - F_P P_i)^2}{N-1} \tag{3}$$

is the residual variance of the differences which, for a large N and from the central limit theorem, can be stated to follow a chi-squared distribution with $N-1$ degrees of freedom [12].
Now, given two different prediction vectors $\mathbf{P_1}$ and $\mathbf{P_2}$ we can compute their corresponding V_1 and V_2 with Eq. 3 and compute the ratio

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

which, by definition, follows the distribution of an F variable [13]. It is easy to see that

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

Using F_{test}, we can now formulate the null hypothesis that $\mathbf{P_1}$ and $\mathbf{P_2}$ have no significant differences in predicting $\mathbf{G}$. This hypothesis must be rejected when $F_{test} < F_C$ or $F_{test} > \frac{1}{F_C}$, where F_C is the critical value of the two-tailed F distribution with a certain (usually 95%) confidence level and $(N-1, N-1)$ degrees of freedom.
Consequently, using F_{test}, we may conclude that predictions $\mathbf{P_1}$ and $\mathbf{P_2}$ are equal ($F_{test} = 1$), insignificantly different ($F_C \leq F_{test} \leq \frac{1}{F_C}$), or significantly different ($F_{test} < F_C$ or $F_{test} > \frac{1}{F_C}$). In the latter case, the one having the lowest value of V would be significantly better than the other.

3 Experimental results

3.1 Synthetic datasets

Aiming to perform synthetic experiments that allow us to characterize the performance of STRESS in front of Pearson and Spearman correlations, we generated a dataset of groundtruth and predicted data using random values. In particular, we have generated 500 pairs of groundtruth and predicted data using a uniformly distributed probability function in the $[0, 5]$ interval. Initially, this random generation would also provide random results for the two correlation measures, and for STRESS. From these data we will study how the agreement measures behave when improving the agreement between groundtruth and prediction data. In addition, we will analyze how the introduction of outliers affects the correlations and STRESS.
Therefore, we started by reducing in an increasing way the difference between the groundtruth and predicted data in each pair by modifying the predicted data towards the groundtruth in a fixed percentage of each difference ($|G_i - P_i|, i \in \{1, ..., N\}$) from 0% to 100% in steps of 10% so that, eventually, we obtain perfect data agreement. We run this experiment five times with different random initial values. In Figure 1 (left) we have plotted the average results and standard deviations (multiplyed by 3 for visualization purposes) between the 5 experiments, provided for STRESS, Pearson and Spearman coefficients. For clarity of presentation, all measures have been re-scaled in the interval $[0, 100]$ as shown in the legend of the Figure 1 (left). It is clear that the curve for STRESS is almost linear while the ones for Pearson and Spearman are highly nonlinear. In particular, it is specially interesting to note that when differences between predicted and groundtruth data have been reduced 80% or more, Pearson and Spearman correlations have little sensitivity in this range, while STRESS has the same sensitivity in every reduction step. This region of good agreement between groundtruth and predicted data is the usual case in most of the applications of these indexes.
Second, starting again with 500 pairs of random values for G and P, and also repeating five times the computations, ramdon outliers are introduced in the generated data. In particular, each outlier corresponds to multiplying by a factor of 10 one random pair prediction, also randomly chosen. That is, we arbitrarily considered an outlier as a change of one order of magnitude. We proceeded by increasingly introducing one by one more outliers from one until a number of 10, which corresponds from 0.2% to 2% of the whole dataset. In Figure 1 (right) we have plotted the relative worsening observed for STRESS, Pearson and Spearman coefficients (computed in the range $[0, 100]$ as commented above), with respect to their initial values, when introducing one by one the outlier while keeping the previous ones in each case. Specifically, we plotted the average and standard deviations (divided by 5, for visualization purposes) of the relative worsening. We

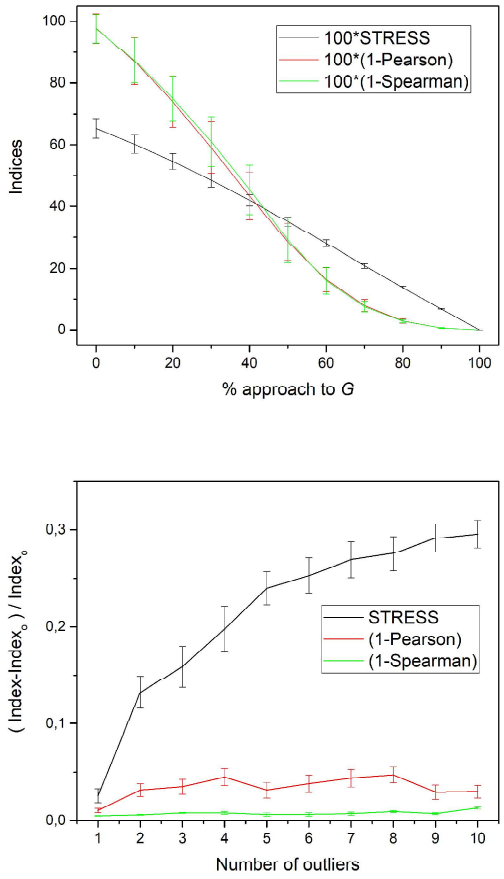

Figure 1: Left: STRESS values, and Pearson and Spearman correlations when increasingly reducing the differences between random groundtruth data and random predictions. Right: Relative worsening of STRESS values, and Pearson and Spearman correlations when increasing the number of outliers in the data set.

can see the the worsening ratio has an up to 30% increment for 10 outliers in the case of STRESS. This plot also shows that STRESS is much more sensitive to the introduction of outliers than Pearson and Spearman correlations, being the latter almost insensitive to them.

3.2 Image quality scores dataset

Now we compare the performance of STRESS with the classical Pearson and Spearman coefficients when predicting image quality scores for a real experimental dataset. As groundtruth data we use the image quality scores dataset in the Colourlab Image Database: Image Quality (CID:IQ) [11]. This dataset contains 23 pictorial images selected as the reference images with 6 different distortions over 5 levels. The distortions are JPEG compression, JPEG2000 compression, Poisson noise, blurring, and two gamut mapping algorithms. These images where evaluated by a total of 17 observers.

In order to predict these data values, we use 10 image quality metrics applied between each distorted image and the corresponding reference: (1) image Color Appearance Model difference (iCAMd) [14], (2) Fuzzy Color Structural Similarity (FCSS) [15], (3) Structural Similarity Index (SSIM) [1], (4) Multiscale Structural Similarity Index (MSSIM) [16], (5) Color Structural Simi-

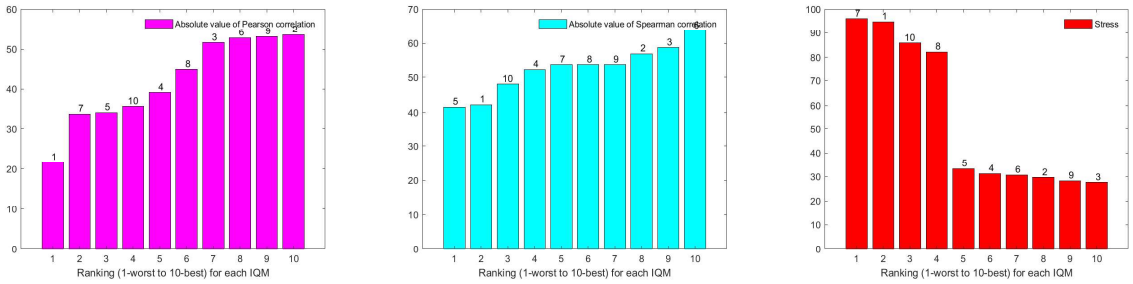

Figure 2: Ranks from worst (1) to best (10) obtained for each of the Image Quality Metrics in the comparison and for each index (rescaled to $[0, 100]$): Pearson correlation (left), Spearman correlation (center) and Stress (right). Absolute values of correlations are used.

IQM	iCAMd	FCSS	SSIM	MSSIM	CSSIM	FSIMc	MSE	RMSE	PSNR	NCD
iCAMd										
FCSS										
SSIM										
MSSIM										
CSSIM										
FSIMc										
MSE										
RMSE										
PSNR										
NCD										

Table 1: Statistical significant performance in terms of STRESS for 95% confidence interval. For each pair of image quality metrics (IQM), we fill the cell with a green colour when performance of the metrics is significantly different and with the red colour, otherwise.

larity Index (CSSIM) [17], (6) Feature similarity index (FSIMc) [18], and the classical ones (7) Mean Squared Error (MSE), (8) Root Mean Squared Error (RMSE), (9) Peak Signal to Noise Ratio (PSNR), and (10) Normalized Color difference (NCD) [19].
In Figures 2 (left to right), we compare the agreement between the image quality metrics and the average observers scores, given by the Pearson and Spearman correlation coefficients and STRESS. Absolute value of correlations found is used as it is meaningless to us whether the correlations are direct or inverse. In each bar plot, the 10 metrics are sorted from left to right corresponding to from worst to best agreement predicted in each case. As we can see, the order differs significantly depending on the agreement measure, but in all cases there are some quality measures with a similar performance. The statistical significance test explained in Section 2.1 can be used in the case of STRESS in order to determine whether the differences are really meaningful or not. Thus, Table 1 shows the statistical significance tests computed for all pairs of metrics. Each position in the double entry table represents whether the two corresponding metrics show a significantly different performance (green colour) or not (red colour) for a 95% confidence level. It is interesting to note that all metrics have a performance that it is not significantly different from at least one other metric. In particular, among the best performing metrics (FCSS, SSIM and PSNR) we found not significant differences for a 95% confidence level.

4 Conclusions

In this work we have studied the application of the standardized residual sum of squares (STRESS) as an alternative to Pearson and Spearman correlation coefficients to measure the agreement between psychophysical groundtruth data of image quality and computed image quality metrics. From synthetic experiments we have seen than STRESS has more sensitivity for smaller differences between predictions and groundtruth data and it is also more sensitive to outliers in the dataset. When applying STRESS to an image quality database, we saw how useful it is to have the possibility to run statistical significance tests to decide whether performance differences among image quality metrics can be considered meaningful or not. In this case, we found that there are no statistically significantly different results between the best performing metrics, so we may wonder if this performance can be really improved.

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 Investigación", Spain) along with the European Union FEDER (European Regional Development Funds) support.

References

[1] Z Wang, AC Bovik, HR Sheikh, EP Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE transactions on image processing 13 (4), 600-612.

[2] `https://www.scopus.com/authid/detail.uri?authorId=56984291600`

[3] J.L. Rodgers, W.A. Nicewander, "Thirteen ways to look at the correlation coefficient", *Am Stat* 42, 59–66, 1988.

[4] C. Spearman, "The Proof and Measurement of Association between Two Things", *American Journal of Psychology* 15, 72–101 (1904).

[5] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Pychometrika* 29, 1–27 (1964).

[6] A. P. M. Coxon, The User's Guide to Multidimensional Scaling (Heinemann, 1982).

[7] P.A. Garcia, R. Huertas, M. Melgosa, G. Cui, "Measurement of the relationship between perceived and computed color differences" in *JOSA - A* 24 (7), 1823-1829, 2007.

[8] C. Li, Z. Li, Z. Wang, Y. Xu, M.R. Luo, G. Cui, M. Melgosa, M.H. Brill, and M. Pointer, Comprehensive color solutions: CAM16, CAT16, and CAM16-UCS, in *Color Research and Applications* 42 (6), 703-718, 2017.

[9] M. Safdar, G. Cui, Y.J. Kim, and M.R. Luo, "Perceptually uniform color space for image signals including high dynamic range and wide gamut," *Opt. Express* 25, 15131-15151 (2017)

[10] M. Pedersen, J.Y. Hardeberg, "Full-Reference Image Quality Metrics: Classification and Evaluation", in *Foundations and Trends® in Computer Graphics and Vision* 7 (1), 1-80, 2012.

[11] X. Liu, M. Pedersen, J.Y. Hardeberg, "CID: IQ–a new image quality database", *International Conference on Image and Signal Processing*, 193-202, 2014.

[12] J.F. Seely, D. Birkes, Y. Lee, "Characterizing Sums of Squares by Their Distributions." The American Statistician, 51 (1), 55–58 (1997). JSTOR `www.jstor.org/stable/2684696` `https://doi.org/10.2307/2684696`

[13] R.A. Fisher, "On the mathematical foundations of theoretical statistics". Philosophycal Transactions of the Royal Society of London - A: Mathemathical, Physical and Engineering Sciences, 222, 309–368 (1922).

[14] M.D. Fairchild, G.M. Johnson, "iCAM framework for image appearance, differences, and quality", *Journal of Electronic Imaging* 13 (1), 126-138.

[15] S. Grečova, S. Morillas, "Perceptual similarity between color images using fuzzy metrics" *Journal of Visual Communication and Image Representation* 34 (2016), 230-235.

[16] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," Invited Paper, IEEE Asilomar Conference on Signals, Systems and Computers, Nov. 2003

[17] M. Hassan, C. Bhagvati, Structural Similarity Measure for Color Images, International Journal of Computer Applications 43(14)(2012)7—12.

[18] Lin Zhang, Lei Zhang, X. Mou, D. Zhang, FSIM: A Feature Similarity Index for Image Quality Assessment, IEEE Trans. Image Processing 20(8)(2011) 2378–2386. Source code is available at `http://www4.comp.polyu.edu.hk/~cslzhang/IQA/FSIM/FSIM.htm`.

[19] K.N. Plataniotis, A.N. Venetsanopoulos, Color Image Processing and Applications, Springer-Verlag, (2000) 355 pp 1-45, 51-100, 109-157.