Correlation Coefficients: Reliable Performance Indices for Visual Quality Assessment?

Tsung-Jung Liu[1], Kuan-Hsien Liu[2] , and Hsin-Hua Liu[3]

[1]*Assistant Professor, Dept. of Electrical Engineering, National Chung Hsing University, Taiwan*

[2]*Assistant Professor, Dept. of Electrical Engineering, Chinese Culture University, Taiwan*

[3]*Dept. of Electrical Engineering, National Taiwan University, Taiwan*

Abstract - For visual quality assessment, correlation coefficients (CCs) have always been used as indices to measure the performances of objective visual quality metrics. However, there is a limitation for these indices. In this paper, we investigate this limitation and choose the width of the confidence interval to measure the reliability of three CCs on several visual quality databases. We are also able to determine the required sample sizes for the corresponding databases. Experimental results demonstrate that different CCs have their respective advantages on different circumstances. For instance, Spearman Rank Order CC is the most reliable index among three CCs when the considered visual quality metrics and subjective quality scores are at high correlation. Furthermore, we also can conclude that a larger sample size is a prerequisite to maintain more reliable CCs.

Keywords: Correlation Coefficient (CC), Confidence Interval (CI), Reliability, Sample Size, Visual Quality Database.

I. INTRODUCTION

Correlation coefficients (CCs) have been adopted to measure the association between the subjective quality scores (mean opinion score (MOS) or differential mean opinion score (DMOS)) and the objective quality scores obtained by visual quality metrics (VQMs) [1, 2]. Some of this type of example are shown in [3-5]. Correlation coefficients also can be used to measure the difference between the quality scores generated by different subjective evaluation methods [6]. However, there is little work to be done to address the accuracy and suitability of the CCs applied in this domain. As we know, in order to simplify the computation of CCs, sample CCs are used instead of population CCs. Due to this reason, the reliability of sample CCs depends on the sample size and the population correlation.

In this paper, we plan to investigate the reliability of CCs (including Pearson's CC (PCC), Spearman's rank order CC (SROCC), and Kendall's rank CC (KRCC), which are the three most commonly used performance indices) in several well-known image and video quality databases, such as LIVE image [7], CSIQ [8], TID2008 [9], TID2013 [10], and LIVE video [11] databases. Also, we propose to use the approach in [12] to find the proper sample size for specified correlation coefficients and target reliability (i.e., acceptable confidence interval (CI) for CCs) for each database. Then, we can have a clue what is the required sample size to build a visual quality database to meet our needs.

The rest of the paper is organized as follows. The three popular correlation coefficients will be described concisely in Section II. In Section III, we show how to calculate the desired widths of confidence intervals for three corresponding CCs. Section IV introduces the steps to determine the required sample size for specified confidence interval and its width. Extensive experiments on five well-known and publicly available visual quality databases are done in Section V. We also present the detailed discussion and comprehensive analysis in Section V. The final concluding remarks are drawn in Section VI.

II. CORRELATION COEFFICIENTS (CCS)

In this section, we will briefly introduce three commonly seen correlation coefficients used in measuring the performance of visual quality metrics [13, 14].

First, Pearson's correlation coefficient (PCC) [15, 16] is the covariance of the two variables divided by the product of their standard deviations. When it is applied to a population, we usually call it as population Pearson correlation coefficient. And it can be represented by

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}, \tag{1}$$

where $Cov(X,Y)$ is the covariance of X and Y, and σ_X, σ_Y represent the standard deviation of X and Y, respectively. (1) can also be expressed by

$$\rho_{X,Y} = \frac{E[(X-E[X])(Y-E[Y])]}{\sqrt{(X-E[X])^2}\sqrt{(Y-E[Y])^2}}, \tag{2}$$

where $E[X]$, $E[Y]$ are the mean of X and Y, respectively. However, when it is applied to a sample, we refer it to as sample Pearson correlation coefficient. Suppose we have two datasets $\{x_1, x_2, \cdots, x_n\}$ and $\{y_1, y_2, \cdots, y_n\}$, then the sample Pearson correlation coefficient is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \quad (3)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample means.

Secondly, the Spearman's rank order correlation coefficient (SROCC) [17] is defined as the Pearson's correlation coefficient between the ranked variables [18]. For a sample of size *n*, the *n* raw scores x_i, y_i are converted to ranks R_{x_i}, R_{y_i}, and the sample SROCC can be computed by

$$\rho = 1 - \frac{6\sum_{i=1}^{n}\left(R_{x_i}-R_{y_i}\right)^2}{n(n^2-1)}. \quad (4)$$

The third correlation coefficient introduced here is the Kendall's rank correlation coefficient (KRCC) [19]. Any pair of (x_i, y_i) and (x_j, y_j), $i, j = 1, 2, \cdots, n$, where $i \neq j$, are said to be concordant if the ranks for both elements agree (i.e., both $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$). They are said to be discordant if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant. The sample KRCC is defined as

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}, \quad (5)$$

where n_c is the number of concordant pairs, and n_d is the number of discordant pairs.

Among the three correlation coefficients mentioned above, PCC is sensitive only to a linear relationship between two variables. The other two correlation coefficients (SROCC and KRCC) belong to rank correlation coefficients, which are developed to be more robust than the PCC (i.e., more sensitive to nonlinear relationships). It is common to consider SROCC and KRCC as alternatives to PCC, which are used to reduce the amount of calculation or to make the coefficient less sensitive to non-normal distributions. However, there is another different point of view [20] saying this lacks mathematical basis since rank correlation coefficients are aimed to measure a different type of relationship than the PCC and are best seen as measures of a different type of association instead of alternative measures of population correlation coefficient.

III. CONFIDENCE INTERVALS FOR CCS

Suppose the population CC is denoted as z_r, and the sample CC is denoted as *r*. Based on the results of Bonett and Wright in [12], if the population is bivariate normally distributed, we can use Fisher transformation (also called *r* to z_r transformation) below

$$z_r = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \quad (6)$$

to convert *r* to z_r, which is approximately normally distributed with variance

$$Var(z_r) = \frac{c}{n-b}, \quad (7)$$

where c = 1, $1 + r^2/2$, 0.437, and b = 3, 3, 4 for Pearson, Spearman and Kendall correlations, respectively. In other words, the distribution of z_r may not be strictly normal, but it would be very like the normal distribution when the sample size increases. Therefore, the upper and lower confidence limits for z_r can be computed by

$$z_u = z_r + z_{(1-\alpha/2)}\sqrt{\frac{c}{n-b}}, \quad (8)$$

$$z_l = z_r - z_{(1-\alpha/2)}\sqrt{\frac{c}{n-b}}, \quad (9)$$

where $z_{(1-\alpha/2)}$ is the 100(1-α/2) percentage point of the standard unit normal distribution. The values of z_u and z_l are then transformed back to the confidence limits of *r* using

$$r_u = tanh(z_u) = \frac{e^{2z_u}-1}{e^{2z_u}+1}, \quad (10)$$

$$r_l = tanh(z_l) = \frac{e^{2z_l}-1}{e^{2z_l}+1}. \quad (11)$$

The desired width for the confidence interval of *r* is

$$w = r_u - r_l. \quad (12)$$

IV. SAMPLE SIZE DETERMINATION

The sample size required to obtain a 100(1- α)% confidence interval with a desired width *w* can be computed by the following two stages.

In the first stage, sample size approximation is calculated by

$$n_0 = \left\lceil 4c(1-r^2)^2\left(\frac{z_{(1-\alpha/2)}}{w}\right)^2 + b \right\rceil, \quad (13)$$

where $\lceil\cdot\rceil$ denotes the ceiling function and we set n_0 = 10 if n_0 < 10.

Let w_0 denote the width of Fisher confidence interval for sample size n_0, and *n* denote the sample size that yields a Fisher confidence interval having desired width *w*. In the second stage approximation, we compute the required sample size *N* via

$$N = \left\lceil (n_0 - b)\left(\frac{w_0}{w}\right)^2 + b \right\rceil. \quad (14)$$

V. EXPERIMENTS AND DISCUSSIONS

TABLE 1. THE SAMPLE SIZE FOR EACH VISUAL QUALITY DATABASE

Database	DB1	DB2	DB3	DB4	DB5
Name	LIVE Image	CSIQ	TID2008	TID2013	LIVE Video
Sample Size	779	866	1700	3000	150

TABLE 2. THE KURTOSIS VALUE FOR DATA IN EACH VISUAL QUALITY DATABASE

Database	DB1	DB2	DB3	DB4	DB5
Kurtosis Coefficient	2.0793	2.1527	3.1380	2.8566	2.4340

TABLE 3. CCS AND WIDTH OF CI FOR THE VQMS IN EACH DATABASE

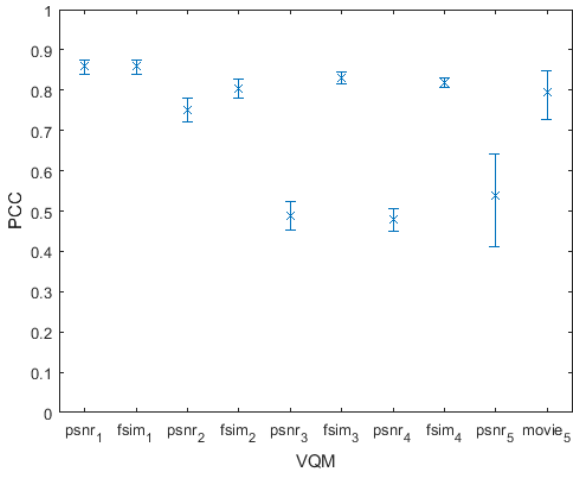
Database	VQM	PCC	width of CI	SROCC	width of CI	KRCC	width of CI
DB1	PSNR	0.8585	**0.0371**	0.8756	0.0387	0.6865	0.0492
	FSIM	0.8586	0.0371	0.9634	**0.0123**	0.8337	0.0284
DB2	PSNR	0.7512	0.0582	0.8057	**0.0540**	0.6078	0.0557
	FSIM	0.8048	0.0471	0.9242	**0.0233**	0.7561	0.0378
DB3	PSNR	0.4890	0.0724	0.5245	0.0736	0.3696	**0.0543**
	FSIM	0.8300	0.0296	0.8805	**0.0252**	0.6946	0.0326
DB4	PSNR	0.4785	0.0552	0.6394	0.0465	0.4696	**0.0369**
	FSIM	0.8195	**0.0235**	0.8015	0.0294	0.6289	0.0286
DB5	PSNR	0.5372	0.2297	0.5205	0.2507	0.3646	**0.1855**
	MOVIE	0.7955	**0.1196**	0.7890	0.1411	0.6019	0.1368

In order to have a comprehensive study of the reliability of CCs on the existing well-known visual quality databases, we use 4 image quality databases (LIVE image [7], CSIQ [8], TID2008 [9], TID2013 [10]) and 1 video quality database (LIVE video [11]) for the test. In each database, the CCs are used to measure the association between the subjective scores (MOS or DMOS) and the objective scores computed by VQMs. For image quality databases, we use PSNR and FSIM [21], which are the baseline and the best full-reference (FR) formula-based VQMs in the field of image quality assessment. Similarly, PSNR and MOVIE [22] represent the baseline and the best FR formula-based VQMs in evaluating the video quality. The details of sample size for each quality database are listed in Table 1.

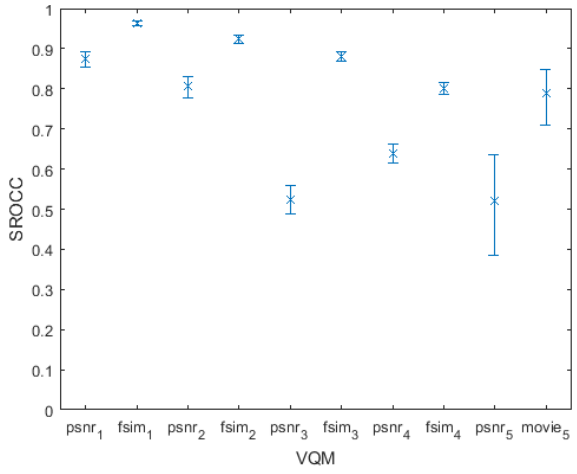
In addition, to apply the formulas introduced in Sections III and IV, the assumption of data bivariate normally distributed has to be verified. According to [23], we can determine whether the distribution of data in each database is normally distributed by checking the kurtosis coefficient:

$$k = \frac{E[(x-\mu)^4]}{\sigma^4}, \quad (15)$$

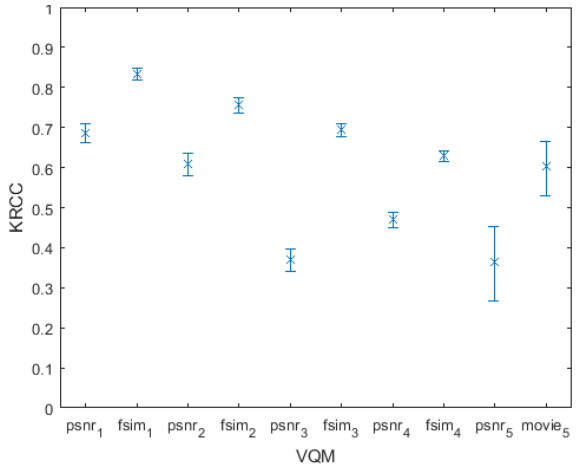
where μ and σ are the mean and standard deviation of x. If the kurtosis coefficient falls between 2 and 4, then the data are considered to form a normal distribution. The kurtosis values for each database are summarized in Table 2. As we can observe from Table 2, the data in each visual quality database are indeed normally distributed because their kurtosis values lie between 2 and 4.



(a)



(b)



(c)

Fig. 1 CCs and the corresponding upper and lower confidence limits for VQMs in each database, where "**x**" denotes the value of CCs, and $xxxx_i$ (i=1-5) represents the use of xxxx (e.g., psnr,

fsim, movie) in database i. (a) PCCs with their upper bound and lower bound of CIs. (b) SROCCs with their upper bound and lower bound of CIs. (c) KRCCs with their upper bound and lower bound of CIs.

In Table 3, we summarize the computed CCs and corresponding width of CIs. FSIM and MOVIE are better VQMs than traditional PSNR since they provide larger correlation with the ground truth (i.e., MOS or DMOS). Moreover, the CCs for FSIM and MOVIE are also more reliable than PSNR because of the smaller width of CIs. Also, by observing Fig.1, we find that both PCC and SROCC have smaller widths on CIs among the three CCs when the CC value is greater than 0.8, which means PCC and SROCC are more reliable under high correlations. On the contrary, KRCC is a more reliable CC than others when the CC has a value smaller than 0.5 (i.e., when being at low correlation).

Suppose in each database, the expected values of SROCC are set at values 0.9634, 0.9242, 0.8805, 0.8015, and 0.7890 for DB1, DB2, DB3, DB4, and DB5, respectively (i.e., the higher SROCC values in Table 3), the required sample size decreases when the width of CIs increases, as shown in Fig. 2. It means that we only need smaller sample size when the reliability of CC is not a strict matter.

TABLE 4. REQUIRED SAMPLE SIZE (N) TO ACHIEVE THE TARGET WIDTH ($w = 0.02$) FOR CIS UNDER EXPECTED SROCC AS INDICATED IN THE TABLE

Database	DB1	DB2	DB3	DB4	DB5
Expected SROCC	0.9634	0.9242	0.8805	0.8015	0.7890
N	298	1174	2698	6494	7183

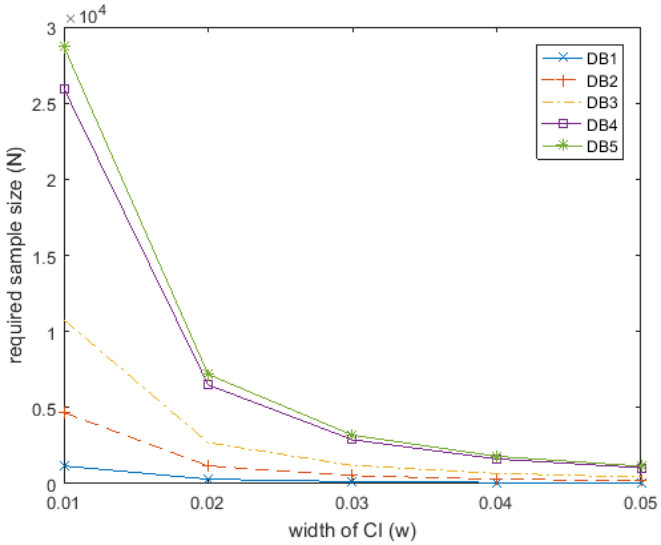


Fig. 2. Required sample size (N) for each database to achieve the target width (w = 0.01 to 0.05) of CIs under fixed SROCC values (DB1: 0.9634, DB2: 0.9242, DB3: 0.8805, DB4: 0.8015, DB5: 0.7890).

Assume that we expect to achieve the value 0.02 for width of CIs, the required sample sizes under expected SROCCs for each database are computed by the steps presented in Sections III, IV and the results are listed in Table 4. Comparing Table 4 with Table 1, the required sample size is much larger than original sample size in each database, except DB1. Especially for video database (DB5), the original sample size is too small to have a reliable CC. Thus, we need to include more image or video samples when building such kind of databases in the future if the reliability of CCs is still a concerned issue.

VI. CONCLUSION AND FUTURE WORK

To the best of our knowledge, the reliability of CCs has not been well and thoroughly discussed in the existing literature. In this work, we use the width of CIs as a way to determine the reliability of CCs. We also discover that each type of CC has its own strength with respect to the specific kind of scenario. In addition, we manage to find the required sample size for several commonly used visual quality databases. The results suggest a larger sample size would be essential for a confident CC.

In the near future, we will try to investigate the width of CIs for each type of correlation coefficients (e.g., PCC, SROCC, and KRCC) and the required sample size for each visual quality database when data are not bivariate normally distributed. As far as we know, this question has not been explored and answered in the existing literature. This will be a very interesting and challenging research topic for us to study.

REFERENCES

[1] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment, phase I,” Mar. 2000. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI.

[2] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II,” Aug. 2003. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII.

[3] T.-J. Liu, K.-H. Liu, and H.-H. Liu, “Temporal information assisted video quality metric for multimedia,” in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pp. 697–701, 2010.

[4] T.-J. Liu, W. Lin, and C.-C. J. Kuo, “Image quality assessment using multi-method fusion,” *Image Processing, IEEE Transactions on*, vol. 22, no. 5, pp. 1793–1807, 2013.

[5] T.-J. Liu, K.-H. Liu, J. Y. Lin, W. Lin, and C.-C. J. Kuo, "A paraboost method to image quality assessment," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. x, no. x, pp. xx–xx, 2016. (In press)

[6] T.-J. Liu, K.-H. Liu, H.-H. Liu, and S.-C. Pei, "Comparison of subjective viewing test methods for image quality assessment," in *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 3155–3159, 2015.

[7] H.R. Sheikh, Z. Wang, L.K. Cormack, and A.C. Bovik, "LIVE Image Quality Assessment Database Release 2," 2006. [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm.

[8] E.C. Larson, and D.M. Chandler, "Categorical Image Quality (CSIQ) Database," 2009. [Online]. Available: http://vision.okstate.edu/csiq.

[9] N. Ponomarenko, "Tampere Image Database 2008 (TID2008), version 1.0," 2008. [Online]. Available: http://www.ponomarenko.info/tid2008.htm.

[10] N. Ponomarenko, "Tampere Image Database 2013 (TID2013), version 1.0," 2013. [Online]. Available: http://www.ponomarenko.info/tid2013.htm.

[11] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, and L.K. Cormack, "LIVE Video Quality Database," 2009. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html.

[12] D.G. Bonett, and T.A. Wright, "Sample size requirements for estimating Pearson, Kendall and Spearman correlations," *Psychometrika*, vol. 65, no. 1, pp. 23–28, 2000.

[13] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Recent Developments and Future Trends in Visual Quality Assessment," *in Proceedings of 2011 Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Submit and Conference*, pp. 1–10, Oct. 2011.

[14] T.-J. Liu, Y.-C. Lin, W. Lin, and C.-C. J. Kuo, "Visual quality assessment: recent developments, coding applications and future trends," *APSIPA Transactions on Signal and Information Processing*, vol. 2, e4, 2013. [Online]. Available: http://journals.cambridge.org/article S204877031300005X

[15] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240-242, 1895.

[16] J. Lee Rodgers, and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no.1, pp. 59-66, 1988.

[17] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 15, no. 1, pp. 72-101, 1904.

[18] J. L. Myers, A. Well, and R. F. Lorch, *Research design and statistical analysis*, Routledge, 2010.

[19] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81-93, 1938.

[20] M. G. Kendall, *Rank correlation methods*, Griffin, 1948.

[21] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *Image Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2378–2386, 2015.

[22] K. Seshadrinathan, and A.C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *Image Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 335–350, 2010.

[23] ITU, "Methodology for the Subjective Assessment of the Quality of Television Pictures," International Telecommunication Union, Recommendation ITU-R BT.500-13, Jan. 2012.