Extended Hybrid Image Similarity – Combined Full-Reference Image Quality Metric Linearly Correlated with Subjective Scores

K. Okarma[1]
[1]*Department of Signal Processing and Multimedia Engineering, Faculty of Electrical Engineering*
West-Pomeranian University of Technology,
Sikorskiego 37, 70-313 Szczecin, Poland
okarma@zut.edu.pl

***Abstract*—One of the most relevant issues in image processing and analysis is a reliable image quality assessment. During last several years numerous metrics have been proposed by various researchers which are much better than traditionally used Mean Squared Error or similar metrics in the aspect of the accordance with human perception of various distortions. Nevertheless, the direct application of such metrics does not provide high correlation with subjective scores because of the required additional nonlinear mapping. Unfortunately, such fitting, typically applied for each image database using the logistic function, leads to different values of parameters for each dataset. As a more universal approach, some nonlinear combinations of various metrics have been proposed recently which do not require any nonlinear mapping. In the paper an extended combined similarity metric is proposed, which provides high prediction accuracy of the image quality with highly linear correlation with subjective scores. The results of extensive tests conducted using the most relevant image quality assessment databases are also presented.**

***Index Terms*—Image quality assessment, image analysis, image similarity.**

I. INTRODUCTION

The role and importance of the image analysis in various applications is still growing. Regardless of the specific problem, the accuracy of detection, recognition or classification based on the image processing and analysis strongly depends on the quality of input images. In many cases such subjective image quality assessment is conducted by the human operator of the system and may be specific for a given application. Nevertheless, some typical image distortions are common and their impact on the results of the image analysis is similar.

Some other important issues are data transmission and visualization which are directly related to image quality assessment and some of image processing methods. In such applications an objective reliable image quality assessment, independent on the human subject but well correlated with subjective quality scores, may be useful for the development of some new algorithms as well as their optimization and verification, especially for color images.

II. CLASSICAL IMAGE QUALITY METRICS

A reliable objective image quality metric should be independent on the image content so different images subjected to the same distortions should give equal results. Due to the effortless usage for the optimization purposes, scalar metrics are preferred, especially with dynamic range from 0 to 1.

The objective image quality assessment schemes may be classified as belonging to three major groups. The first one is known as “blind” (no-reference) approach, which does not require the access to the original undistorted image at all [1], [2]. Although the potential usage is wide and such metrics are the most desired ones, their universality is currently rather low as they are usually sensitive to only one or two types of distortions. Some typical examples are blur metrics and measures of JPEG artifacts observed as blocks.

Another, more popular group of methods consists of numerous full-reference metrics, with classical Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR). Such methods require the exact knowledge of the original, perfect quality image without any distortions and the quality score is calculated by comparing some features between the distorted and original (reference) images. Although in practical applications the access to the original image is not always possible, a great progress in this family of metrics has taken place in recent years.

The third, less popular approach is known as reduced-reference and it requires a partial knowledge about the reference image e.g. some DCT coefficients or specified features. Such methods may also be used for the estimation of some full-reference metrics [3].

Regardless of the type of the metric, each newly developed one should be verified in order to determine its concordance with subjective evaluations and Human Visual System (HVS) which can be modeled using many techniques describing the way that human observers perceive various kinds of image distortions. For this reason some image quality assessment databases have been developed containing reference and distorted images together with Mean Opinion Scores (MOS) or Differential Mean Opinion Scores (DMOS) collected during the experiments conducted

Manuscript received February 24, 2013; accepted July 18, 2013.

in cooperation with numerous observers.

III. RECENT FULL-REFERENCE IMAGE QUALITY METRICS

Poor correlation of some traditional metrics based on the comparison of corresponding pixels from the reference and distorted images, such as MSE or PSNR, caused the necessity to develop a new approach based on alternative assumptions. The first such idea, known as Universal Image Quality Index, is based on the comparison of three local features corresponding to common distortions, using the sliding window approach [4]. These features are: luminance distortions, loss of contrast and structural distortions. After some modifications, e.g. increasing its stability, the UIQI metric, has been extended [5] into one of the most popular image quality metrics called Structural Similarity (SSIM). The local SSIM formula for each window position can be expressed as the combination of the mean values, variances and the covariance

$$SSIM(x,y)=\frac{\left(2\overline{xy}+C_1\right)\cdot\left(2\pm_{xy}+C_2\right)}{\left(\overline{x}^2+\overline{y}^2+C_1\right)\cdot\left(\pm_x^2+\pm_y^2+C_2\right)}, \qquad (1)$$

assuming that x and y denote the local fragments of distorted and reference images inside 11×11 pixels windows (weighted using Gaussian function) with small stabilizing constants C_1 = $(0.01{\cdot}L)^2$ and C_2 = $(0.03{\cdot}L)^2$ where L is the number of available luminance levels (typically 256).

Due to the popularity of the SSIM metric some modifications have also been proposed e.g. three-component weighted SSIM, gradient SSIM or Multi-Scale SSIM [6], which is defined as

$$MS-SSIM(x,y)=\left[l(x,y)\right]^{\bar{}M}\times$$
$$\times\prod_{j=1}^{M}\left[c(x,y)\right]^{\;j}\cdot\left[s(x,y)\right]^{¥j}. \qquad (2)$$

where components representing the luminance, contrast and structural distortions respectively, are weighted for each j-th scale (but the luminance changes are considered only for full resolution images). The default values of coefficients α, β, and γ have been proposed by the authors of the paper [6] as the result of the optimization procedure with 600 images used for testing involving 8 observers.

Some other interesting ideas of full-reference image quality assessment are based on the applications of Singular Value Decomposition (SVD), wavelets, other transforms or using some elements of the information theory. An example of this approach is the Visual Information Fidelity (VIF) defined as the mutual information that vision extracts from the distorted image divided by the information extracted from the reference one, calculated for several sub-bands in the wavelet domain [7].

One of the most promising directions of research seems to be the similarity based approach represented by Riesz-based Feature Similarity (RFSIM) and Feature Similarity (FSIM), which are quite similar to the idea of the SSIM. The RFSIM metric [8] is defined as

$$RFSIM=\prod_{i=1}^{5}\frac{\sum_u\sum_v d_i(u,v)\cdot M(u,v)}{\sum_u\sum_v M(u,v)}, \qquad (3)$$

where M is the binary mask being the edge detection result (for this purpose well-known Canny filter may be applied) and d_i are local similarity values calculated for five features obtained using the 1st and 2nd order Riesz transform coefficients. The local similarity for the images x and y can be calculated (using small stabilizing constant value C) as

$$d_i(u,v)=\frac{2\cdot x_i(u,v)\cdot y_i(u,v)+C}{x_i^2(u,v)+y_i^2(u,v)+C}. \qquad (4)$$

Further research of the same group has led to the definition of the FSIM metric [9] based on two factors: phase congruency (PC) and gradient magnitude (G). The construction of the overall index is quite similar as in (3) and its value can be obtained as

$$FSIM=\frac{\sum_u\sum_v S(u,v)\cdot PC_{\max}(u,v)}{\sum_u\sum_v PC_{\max}(u,v)}, \qquad (5)$$

where PC_{max} is the higher from two local values of phase congruency from the reference and assessed image. The local similarity value is defined as the product of two factors related to gradient (Scharr filter is recommended for this purpose) and phase congruency

$$S(u,v)=\left(\frac{2\cdot PC_1(u,v)\cdot PC_2(u,v)+T_{PC}}{PC_1^2(u,v)+PC_2^2(u,v)+T_{PC}}\right)^{\sim}\times$$
$$\times\left(\frac{2\cdot G_1(u,v)\cdot G_2(u,v)+T_G}{G_1^2(u,v)+G_2^2(u,v)+T_G}\right) . \qquad (6)$$

The color version of the metric can also be calculated using the YIQ color model in a similar way, replacing the PC and G values by the chrominance I and Q respectively and multiplying the obtained result (using the exponent value ¥=0.03) by the formula (6), assuming the values of exponents ~ and   equal to 1.

All the metrics briefly presented above have an important common disadvantage – their values are not directly related to the subjective scores expressed as MOS or DMOS values so the additional nonlinear mapping is required in order to achieve high values of the linear correlation coefficients between objective and subjective quality scores.

IV. IMAGE QUALITY ASSESSMENT DATASETS

The verification of the compliance of objective metrics with MOS or DMOS values can be expressed as the quality prediction accuracy and prediction monotonicity. The accuracy is measured using Pearson's linear Correlation Coefficient (CC) whereas the monotonicity can be evaluated by Spearman Rank Order Correlation Coefficient (SROCC) or Kendall Rank Order Correlation Coefficient (KROCC).

In order to achieve a reliable verification, calculations should be conducted for all available datasets. The most relevant of them is Tampere Image Database [10] containing 1700 color images with 17 types of distortions with MOS values obtained from 838 observers. The other two relevant datasets are Categorical Subjective Image Quality (CSIQ) database from Oklahoma State University [11] with 866 images (35 observers and 6 distortion types) and well-known LIVE dataset from Texas University at Austin [12] containing 779 images distorted in five ways assessed by 29 subjects.

As a supplement for those three databases, less important ones may also be used such as: IRCCyN/IVC [13] from University of Nantes (160 images with 4 types of distortions assessed by 15 observers), Wireless Image Quality (WIQ) with 80 distorted greyscale images judged by 30 observers [14] and A57 dataset [15] consisting of 54 test images with 6 types of contaminations evaluated by 7 experts. The oldest database has been developed by Toyama University in Japan and is known as MICT database [16]. Nevertheless, its usefulness is currently strongly limited as it contains 198 images assessed by 16 students but only two types of distortions related to JPEG and JPEG2000 compression.

V. Proposed Approach and the Verification Results

Highly linear relationship between the objective and subjective scores is typically obtained by nonlinear mapping, e.g. by logistic function, with necessary optimization of the mapping function’s parameters. Unfortunately it ought to be conducted independently for each dataset leading to different values of the coefficients. For this reason such approach cannot be considered as a universal one.

Much better results can be obtained using the nonlinear combination of some metrics as proposed in the paper [17]. Using the weighted product of three (or more) metrics with exponent values optimized for the largest database (TID) a serious increase of the CC values can be achieved in comparison to each of the metrics separately (even after nonlinear mapping), e.g. the combination of MS-SSIM, VIF and R-SVD metrics leads to CC = 0.86 [17] and replacing the R-SVD by FSIMc metric proposed in the paper [18] as CISI metric leads to CC = 0.8752 for the TID database.

Good results may also be obtained using the nonlinear combination of the RFSIM and FSIMc metrics leading to HFSIMc metric with CC = 0.8861 for the same dataset.

Another possibility is changing the weighting exponents inside the calculation procedure of the FSIM – ~ and in (6) – or FSIMc metric, discussed in [22], leading to the Weighted FSIM (WFSIM) metric and its color version WFSIMc, increasing the values of the rank order correlation coefficients with subjective scores.

The extended version of the approach based on the combination of four metrics: MS-SSIM, VIF, RFSIM and weighted FSIM is proposed in this paper defined as

$$EHIS = \left(MS - SSIM\right)^{a} \times \left(VIF\right)^{b} \times \left(WFSIMc\right)^{c} \times \left(RFSIM\right)^{d}, \quad (7)$$

assuming using the color version of the WFSIM metric for available color images. The results of the exponents obtained as the result of optimization conducted using TID database are

$$\begin{bmatrix} a & b & c & d \end{bmatrix} = \begin{bmatrix} -1.6131 & 0.2037 & 59.7151 & 0.1989 \end{bmatrix}, \quad (8)$$

with the definition of the WFSIM or WFSIMc metric using the values of the coefficients suggested in the paper [19]

$$\begin{bmatrix} \sim \\ \ \end{bmatrix} = \begin{bmatrix} 0.01 \\ 0.05 \end{bmatrix} \quad (9)$$

or

$$\begin{bmatrix} \sim \\ \ \\ ¥ \end{bmatrix} = \begin{bmatrix} 0.01 \\ 0.05 \\ 0.004 \end{bmatrix}. \quad (10)$$

The obtained CC, SROCC and KROCC values and their comparison with other metrics for the databases described in Section III are presented in Tables I-III and the aggregate values for all datasets weighted according to the number of test images in respective datasets are shown in Table IV.

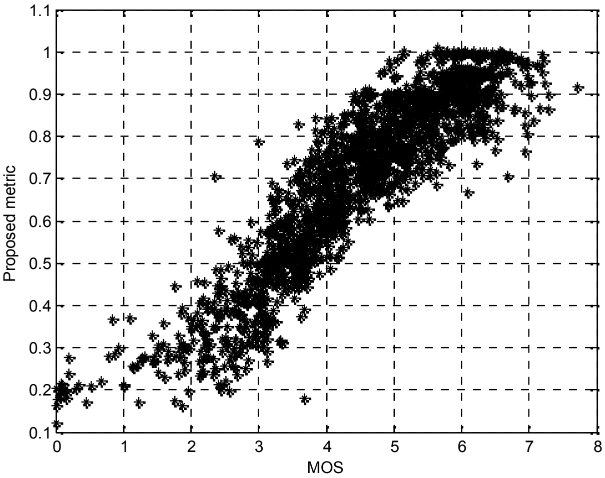

Fig. 1. Scatter plot of the proposed metric versus MOS values for TID2008 database illustrating highly linear relationship between objective and subjective evaluations.

TABLE I. PEARSON LINEAR CORRELATION COEFFICIENTS (CC) FOR VARIOUS METRICS AND DATASETS.

Metric	TID	CSIQ	LIVE	IVC	WIQ	A57	MICT
MS-SSIM	0.7843	0.7708	0.4762	0.7679	0.6089	0.8289	0.7438
VIF	0.7777	0.9219	0.7327	0.8800	0.7301	0.6141	**0.9024**
RFSIM	0.8596	0.9130	0.9352	0.7927	0.7589	0.8419	0.7523
FSIM	0.8300	0.8048	0.8586	0.8563	0.7371	**0.9252**	0.8003
FSIMc	0.8341	0.8208	0.8595	0.8606	---	---	0.8050
HFSIM	0.8853	0.9158	**0.9538**	0.8711	0.7876	**0.9319**	0.7977
HFSIMc	0.8861	0.9197	**0.9532**	0.8721	---	---	0.7992
CISI	0.8752	0.9346	0.9453	**0.9152**	0.7998	0.8663	0.8834
Proposed	**0.9105**	**0.9397**	0.9480	0.8963	**0.8221**	0.7748	0.8940

TABLE II. SPEARMAN RANK-ORDER CORRELATION (SROCC) FOR VARIOUS METRICS AND DATASETS.

Metric	TID	CSIQ	LIVE	IVC	WIQ	A57	MICT
MS-SSIM	0.8526	0.9136	0.9155	0.8845	0.7360	0.8397	0.8864
VIF	0.7496	0.9194	0.7942	0.8964	0.6918	0.6223	0.9086
RFSIM	0.8680	0.9295	0.9401	0.8192	0.7368	0.8215	0.7731
FSIM	0.8805	0.9242	**0.9634**	**0.9262**	0.8006	**0.9181**	0.9059
FSIMc	0.8840	0.9310	**0.9645**	**0.9293**	---	---	0.9067
HFSIM	0.8911	0.9406	0.9605	0.8898	0.7858	**0.9250**	0.8430
HFSIMc	0.8925	0.9422	0.9604	0.8908	---	---	0.8437
CISI	0.8742	0.9426	0.9618	0.9201	0.7845	0.8709	**0.9106**
Proposed	**0.9098**	**0.9498**	0.9622	0.9076	**0.8266**	0.9177	0.9045

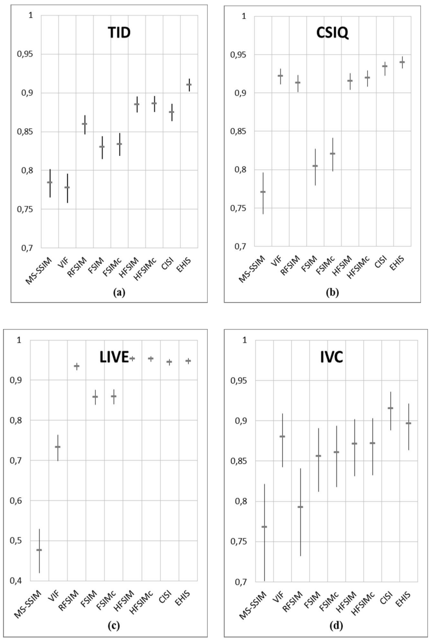

Fig. 2. Lower and upper bounds for the 95% confidence interval calculated for Pearson correlation with subjective scores for major datasets.

TABLE III. KENDALL RANK-ORDER CORRELATION (KROCC) FOR VARIOUS METRICS AND DATASETS.

Metric	TID	CSIQ	LIVE	IVC	WIQ	A57	MICT
MS-SSIM	0.6543	0.7389	0.7431	0.7005	0.5645	0.6483	0.7029
VIF	0.5863	0.7532	0.5848	0.7158	0.5246	0.4594	0.7329
RFSIM	0.6780	0.7645	0.7816	0.6452	0.5493	0.6324	0.5752
FSIM	0.6946	0.7567	**0.8337**	**0.7564**	0.6215	**0.7639**	0.7302
FSIMc	0.6991	0.7690	**0.8363**	**0.7636**	---	---	0.7303
HFSIM	0.7108	0.7900	0.8254	0.7162	0.6038	**0.7667**	0.6479
HFSIMc	0.7125	0.7931	0.8248	0.7192	---	---	0.6485
CISI	0.6896	0.7893	0.8280	0.7488	0.6038	0.6762	**0.7361**
Proposed	**0.7382**	**0.8033**	0.8288	0.7357	**0.6487**	0.7601	0.7260

All the calculations have been conducted using Matlab with Image Processing Toolbox, using also such useful functions as *corrcoef*, as well as *fminsearch* and *fminunc* for optimization purposes.

TABLE IV. CORRELATION COEFFICIENTS FOR VARIOUS METRICS WEIGHTED FOR ALL DATASETS.

Metric	CC	SROCC	KROCC
MS-SSIM	0.7129	0.8796	0.6939
VIF	0.8085	0.8083	0.6336
RFSIM	0.8763	0.8862	0.7086
FSIM	0.8291	0.9093	0.7407
FSIMc	0.8067	0.8831	0.7226
HFSIM	0.8996	0.9121	0.7475
HFSIMc	0.8713	0.8837	0.7256
CISI	0.9032	0.9093	0.7431
Proposed (EHIS)	**0.9195**	**0.9275**	**0.7690**

VI. CONCLUSIONS

Analyzing the results presented in Table IV and Fig. 1–Fig. 2 the advantages of the proposed hybrid similarity metric can be easily noticed for all measures of correlation with subjective evaluations, leading also to the narrower bounds obtained for the 95 % confidence level for Pearson's correlation, especially for three largest datasets.

Proposed metric can be successfully applied for the direct assessment of results of many image processing algorithms e.g. related to compression, filtering or reconstruction. Due to its highly linear correlation with subjective quality evaluations, its application does not require any additional mapping operations leading directly to accurate prediction of image quality regardless of the distortion type.

REFERENCES

[1] X. Li, "Blind image quality assessment", in *Proc. Int. Conf. Image Processing ICIP*, Rochester, New York, 2002, pp. 449–452.

[2] M. Saad, A. Bovik, C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain", *IEEE Trans. Image Processing*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012. [Online]. Available: http://dx.doi.org/10.1109/TIP.2012.2191563

[3] A. Rehman, Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation", *IEEE Trans. Image Processing*, vol. 21, no. 8, pp. 3378–3389, Aug. 2012. [Online]. Available: http://dx.doi.org/10.1109/TIP.2012.2197011

[4] Z. Wang, A. Bovik, "A universal image quality index", *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, Mar. 2002. [Online]. Available: http://dx.doi.org/10.1109/97.995823

[5] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, "Image quality assessment: From error measurement to structural similarity", *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004. [Online]. Available: http://dx.doi.org/10.1109/TIP.2003.819861

[6] Z. Wang, E. Simoncelli, A. Bovik, "Multi-Scale Structural Similarity for image quality assessment", in *Proc. 37th IEEE Asilomar Conf. Signals, Systems and Computers*, vol. 2, Pacific Grove, California, 2003, pp. 1398–1402.

[7] H. Sheikh, A. Bovik, G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics", *IEEE Trans. on Image Processing*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005. [Online]. Available: http://dx.doi.org/10.1109/TIP.2005.859389

[8] L. Zhang, L. Zhang, X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms", in *Proc. 17th IEEE Int. Conf. on Image Processing ICIP*, Hong Kong, China, 2010, pp. 321–324.

[9] L. Zhang, L. Zhang, X. Mou, D. Zhang, "FSIM: A feature similarity index for image quality assessment", *IEEE Trans. Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011. [Online]. Available: http://dx.doi.org/10.1109/TIP.2011.2109730

[10] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, "TID2008 – a database for evaluation of full-reference visual quality assessment metrics", *Advances of Modern Radioelectronics*, vol. 10, pp. 3045, Oct. 2009.

[11] E. Larson, D. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy", *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011006, Jan. 2010. [Online]. Available: http://dx.doi.org/10.1117/1.3267105

[12] H. Sheikh, M. Sabir, A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006. [Online]. Available: http://dx.doi.org/10.1109/TIP.2006.881959

[13] A. Ninassi, P. Le Callet, F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding", in *Proc. SPIE –Human Vision and Electronic Imaging*, no. 6057, San Jose, California, Jan. 2006, pp. 6057–08.

[14] U. Engelke, H.-J. Zepernick, T. Kusuma, "Subjective quality assessment for wireless image communication: The Wireless Imaging Quality database", in *Proc. 5th Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, Arizona, 2010.

[15] D. Chandler, S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images", *IEEE Trans. Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007. [Online]. Available: http://dx.doi.org/10.1109/TIP.2007.901820

[16] Z. Parvez Sazzad, Y. Kawayoke, Y. Horita. (2000) MICT/Toyama image quality evaluation database [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html

[17] K. Okarma, "Combined full-reference image quality metric linearly correlated with subjective assessment", in *ICAISC (1)*, ser. Lecture Notes in Computer Science, L. Rutkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., vol. 6113. Springer, 2010, pp. 539–546.

[18] K. Okarma, "Combined image similarity index", *Optical Review*, vol. 19, no. 5., pp. 249–254, Sep. 2012. [Online]. Available: http://dx.doi.org/10.1007/s10043-012-0055-1

[19] K. Okarma, "Weighted Feature Similarity – a nonlinear combination of gradient and phase congruency for full-reference image quality assessment", in *Image Processing and Communications Challenges 4*, ser. Advances in Intelligent Systems and Computing, R. S. Choraś, Ed., vol. 184. Springer, 2013, pp. 187–194.