On the optimisation of nonlinear mapping functions towards high correlation of full-reference image quality metrics and their combinations with subjective evaluations

Krzysztof Okarma
West Pomerania University of Technology
71–126 Szczecin, ul. 26. Kwietnia 10, e-mail: okarma@zut.edu.pl

One of the most relevant problems related to most of the objective full-reference image and video quality assessment methods is their correlation with subjective perception of various types of distortions. Since the relation between the subjective scores, typically expressed as Mean Opinion Scores (MOS) or Differential MOS, and most of the metrics is nonlinear, various mapping functions are used in order to linearize this relation e.g. logistic function recommended by the Video Quality Experts Group (VQEG). Such compensation can also be conducted using some other function e.g. polynomial or exponential ones. Nevertheless, the results obtained for each of available datasets differ and such approach cannot be considered as universal. One of possible solutions, proposed in some earlier papers, is the use of combined metrics, which have significantly higher linear correlation with MOS or DMOS values without the necessity of nonlinear mapping. The comparative analysis of the optimisation results for some "state-of-the-art" metrics and proposed combined ones is also provided in the paper.

1. Objective image quality assessment methods versus subjective scores

The influence of image quality on the results of many image and video analysis algorithms is one of the most relevant aspects of contemporary image processing and analysis. Since the number of potential application areas of vision based systems is still growing, the challenges related to reliable image quality assessment are still up-to-date. Some examples of such areas are robot vision, automation, mechatronics, biomedical science or Intelligent Transportation Systems (ITS).

During recent 10 years many automatic image quality assessment methods have been proposed differing from each other considering both their types and usefulness. The first group of metrics is based on the subjective evaluation of image quality. Their application for each assessed image requires the feedback from the observers so they cannot be used directly in many applications, such as development of new image processing algorithms (e.g. filtering or lossy compression). Nevertheless, the Mean Opinion Scores (MOS) or Differential MOS (DMOS) values obtained during conducted subjective experiments can be utilised for development of some better objective metrics.

The objective metrics can be divided into three major families: full-reference, reduced-reference and no-reference ("blind") methods. The differences among them are related to the utilisation of the reference image without any distortions having an ideal quality. In many applications such image is unavailable, so the most desired direction of research seems to be the development of no-reference metrics. Nevertheless, such metrics are still not universal as they are sensitive only to one or two types of distortions.

In some applications there is an access to partial information about the reference image and the reduced-reference approach may be applied. However, a great majority of automatic image quality metrics with high universality belongs to the full-reference family.

For some applications classical metrics based on Mean Squared Error (MSE), such as Peak Signal-to-Noise Ratio (PSNR) are sufficient, but in recent years some better methods have been proposed, much better correlated with the Human Visual System (HVS) and therefore much more useful e.g. for the visualisation purposes.

Since there is a need to verify the results obtained using the objective image quality assessment methods, some perceptual experiments have been performed by various researchers from various countries. Their results have been provided as image quality assessment databases containing numerous images with various distortions and the Mean Opinion Scores (MOS) or Differential MOS values useful for the verification of newly proposed objective metrics.

The most relevant available image quality databases are known as LIVE [12], TID2008 [10] and CSIQ [2], although there are also some others. Three major datasets, mentioned above, have been used in this paper for the illustration of the nonlinear mapping problem being its main topic.

The first database has been provided in 2005 by the Laboratory for Image and Video Engineering (LIVE) at Texas University and is probably the most popular one. It contains 5 types of distortions and 779 test images obtained using the 29 originals. Nevertheless, the results obtained using this database are not always representative, so in recent papers some additional verification using newer datasets is usually provided. It can be easily noticed that some of the metrics proposed earlier by various researchers have been "tuned" specifically for the LIVE database as the only available with reasonable number of test images.

Another relevant database of similar size has been provided in 2009 by the research group from Oklahoma State University. This dataset is known as Categorical Subjective Image Quality (CSIQ) and contains 866 images based on 30 reference images corrupted by 6 types of distortions. These images have been assessed by 35 observers using the linear displacement method.

The largest database has been developed at Tampere University in 2008, known as Tampere Image Database (TID2008), which contains 1700 images with 17 types of distortions assessed by totally 838 persons from three countries (Italy,

Finland and Ukraine) using pairwise sorting approach. One of its advantages is the presence of colour distortions, which are not considered in some other databases. Considering its size, it has become probably the most popular during recent two or three years.

Utilising the MOS or DMOS values for each distorted image form the datasets described above it is possible to verify the coincidence of the objective metric's values with subjective scores. The most typical approach for such verification is its comparison with subjective scores taken from such databases by means of the correlation. Three typical methods are applications of Pearson, Spearman or Kendall correlation. However, only the first one is related to the prediction accuracy, therefore it can be treated as the most important one. Pearson linear correlation coefficient (PCC) can be expressed as:

$$PCC(X,Y) = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\cdot\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \cdot \sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}} \tag{1}$$

where X_i and Y_i represent the consecutive values of the subjective and objective image quality metrics.

The application of Spearman Rank Order Correlation Coefficient (SROCC) and similar Kendall correlation (KROCC) is related only to the estimation of prediction monotonicity. Both correlation coefficients are based on the position of both images in two sorted lists according to the objective and subjective quality metrics.

2. Nonlinear mapping between subjective and objective metrics

A perfect objective image quality assessment metric should be correlated with subjective quality scores as well as possible. It should be noted that the correspondence between the objective and subjective scores should be linear leading to high accuracy of predicted quality using the objective metric. Nevertheless, the quality processing in the HVS is nonlinear, so the high values of the PCC for raw values of the metric are hard to obtain. Considering the nonlinear relationship between image distortions and their perceived quality, some additional mapping functions are usually applied to the quality metrics. The most typical approach is the use of the logistic function according to the recommendations [13, 14] of the Video Quality Experts Group (VQEG). However, some researchers apply some other functions, such as polynomials and exponential functions [1] instead of the logistic one.

The nonlinear mapping functions used in this paper are the three-parameter logistic function expressed as:

$$MOS = \frac{b_1}{1+\exp\left(-b_2 \cdot \left(Q - b_3\right)\right)} \tag{2}$$

and five-parameter logistic one:

$$MOS = b_1 \cdot \left(\frac{1}{2} - \frac{1}{1+\exp\left(b_2 \cdot \left(Q - b_3\right)\right)}\right) + b_4 \cdot Q + b_5 \tag{3}$$

where Q stands for the objective metric value and MOS (or DMOS) denotes the subjective score.

In order to verify the necessity of using the logistic function and compare the results, third order polynomial and the 6-parameter exponential function have been also used in the experiments. The 6-parameter exponential function is expressed as:

$$MOS = a_1 \cdot \exp\left(b_1 \cdot Q\right) + a_2 \cdot \exp\left(b_2 \cdot Q\right) + a_3 \cdot \exp\left(b_3 \cdot Q\right) \tag{4}$$

The use of the mapping functions leads to different optimal parameters for different datasets, therefore such approach cannot be considered as universal. Another approach, proposed in author's papers [4-6], is the use of the nonlinear combination of some metrics, significantly increasing the linear correlation with subjective scores. The modified combined metrics proposed in some recent papers [7-9] lead to even better results and will also be discussed in the further part of the paper.

3. State of the art image quality assessment metrics

The verification of the necessity of the nonlinear mapping and its impact on the obtained results has been performed using the most popular modern image quality metrics and the most interesting of the recently proposed ones as well as some combined metrics proposed in author's earlier papers.

Considering the poor correlation of older image quality metrics such as Mean Squared Error (MSE) or Peak Signal to Noise Ratio (PSNR) and the necessity of developing a new approach, the idea of the Universal Image Quality Index (UQI) [15] has been presented in 2002. Its extension known as the Structural Similarity (SSIM), has become probably the most popular image quality metric in recent years inspiring also many other researchers. It is defined as the mean value of the local quality indexes expressed as [16]:

$$SSIM(x,y) = \frac{\left(2\overline{xy} + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\overline{x}^2 + \overline{y}^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{5}$$

with the use of the sliding window approach, assuming 11×11 pixels Gaussian window as the default one. The symbols x and y denote the corresponding fragments of distorted and reference image respectively, whereas C_1 and C_2 are the stability constants preventing the division by zero. The local similarity value is

calculated using the mean values, variances and the covariance so the SSIM metric is sensitive to three types of distortions: luminance, contrast and structural (the three-component form of the metric can be found in the paper [16]). The multi-scale extension of this metric (MS-SSIM), proposed by the same group of researchers, is defined as [17]:

$$MS-SSIM(x,y)=\left[l(x,y)\right]^{\alpha_M}\cdot\prod_{j=1}^{M}\left[c(x,y)\right]^{\beta_j}\cdot\left[s(x,y)\right]^{\gamma_j} \tag{6}$$

assuming the default weighting values for each j-th scale as proposed in the paper [17] being the result of the optimisations conducted by the inventors of this metric using the 600 test images assessed by 8 human observers. The l, c and s symbols denote the luminance, contrast and structural part of the three-component form of the SSIM metric respectively.

An interesting approach to image quality assessment is based on the information theory which has been applied in the Visual Information Fidelity (VIF) metric [11]. This idea is based on the computation of the relative mutual information that vision extracts from the reference and distorted images in the wavelet domain.

Among various ideas applied for image quality assessment some application of the Singular Value Decomposition (SVD) may also be found, leading to promising results. In this case the SVD is applied to the original and the reference images. In one of the most recent metrics based on the SVD the comparison of the singular values and the left singular matrices leads to the definition of the R-SVD metric [3] expressed as:

$$R\text{-}SVD=\sqrt{\frac{\sum_{i=1}^{N_k}(d_i-1)^2}{\sum_{i=1}^{N_k}(d_i+1)^2}} \tag{7}$$

where d_i is the i-th singular value of the referee matrix calculated using the combination of the left singular matrix and the singular values of the distorted image and the right singular matrix of the original one (the detailed description can be found in the paper [3]).

One of the most recent ideas, leading to relatively high correlation of obtained quality scores with subjective evaluations is the Feature Similarity (FSIM) index [18]. This metric is based on the calculation of the phase congruency and gradient map, which can be efficiently obtained using Scharr filter recommended by the authors of the FSIM metric instead of well-known Prewitt or Sobel masks. The local value of the similarity index can be determined according to the following formula:

$$S(x,y)=\left(\frac{2\cdot PC_1(x,y)\cdot PC_2(x,y)+T_{PC}}{PC_1^2(x,y)+PC_2^2(x,y)+T_{PC}}\right)^{\alpha}\cdot\left(\frac{2\cdot G_1(x,y)\cdot G_2(x,y)+T_G}{G_1^2(x,y)+G_2^2(x,y)+T_G}\right)^{\beta} \quad (8)$$

where PC and G stand for the phase congruency and gradient values calculated for both reference and distorted images. The typical weights proposed in the paper [18] are $\alpha = \beta = 1$ and the stabilizing constants T prevent the division by zero, similarly as in the UQI and SSIM metrics. The final value of the FSIM metric can be obtained as:

$$FSIM=\frac{\sum_x\sum_y S(x,y)\cdot PC_m(x,y)}{\sum_x\sum_y PC_m(x,y)} \quad (9)$$

where $PC_m(x,y) = max(PC_1(x,y), PC_2(x,y))$.

Nevertheless, assuming different values of the α and β weighting coefficients even better results can be obtained, leading to the Weighted Feature Similarity (Weight FSIM), proposed in one of the recent papers [8], which can also be applied in the colour version, similarly as the colour Feature Similarity (FSIMc).

A similar idea based on the application of Riesz transform has been proposed earlier by the FSIM inventors in the paper [19]. It is based on the assumption that the most important regions from the perceptual point of view are located near the edges of the objects. For this reason the usage of the mask obtained as the result of using Canny filter with additional dilation are proposed. For such masked image the first and second order Riesz transform coefficients are calculated which are considered as five masked image features. The local similarity of such maps for two images denoted as f and g is conducted in a similar way as for FSIM metric, according to the following formula:

$$d_i(x,y)=\frac{2\cdot f_1(x,y)\cdot g_2(x,y)+C}{f_1^2(x,y)+g_2^2(x,y)+C} \quad (10)$$

for each feature i = 1..5 using the stabilizing constant C. Then, the final RFSIM value can be calculated as:

$$RFSIM=\prod_{i=1}^{5}\frac{\sum_x\sum_y d_i(x,y)\cdot M(x,y)}{\sum_x\sum_y M(x,y)} \quad (11)$$

using the binary mask M obtained after edge filtering.

In order to avoid the necessity of the nonlinear mapping, discussed above, the idea of the Combined Quality Metric (CQM) has been proposed in one of the earlier papers [4], which is in fact the nonlinear combination of three metrics defined as:

$$CQM=(MS-SSIM)^7\cdot(VIF)^{0.3}\cdot(R-SVD)^{-0.15} \quad (12)$$

The exponent values are nearly optimal by means of the Pearson linear correlation coefficient, and have been obtained as the result of optimisation conducted using TID2008 database as the largest currently available one.

This approach has been extended recently leading to the idea of the Combined Image Similarity Metric [7] obtained by replacing the R-SVD metric by the FSIMc and conducting the optimisation of the exponents in the same way as for the CQM. The obtained metric can be expressed as:

$$CISI = (MS - SSIM)^{0.5} \cdot (VIF)^{0.3} \cdot (FSIMc)^{5} \tag{13}$$

A high correlation between the MOS/DMOS values and an objective metric can also be obtained using only two metrics. Such approach has been verified using the combination of FSIM and RFSIM metric leading to the Hybrid Feature Similarity [9]. The optimised values of the exponents for each metric (assuming using the colour version of Feature Similarity) are: 0.4 for RFSIM and 3.5 for FSIMc.

4. Analysis of experimental results

During recent years many new image quality metrics have been delivered by various researchers but most of them are based on the assumption that the additional nonlinear mapping is necessary for obtaining high correlation with subjective scores. The verification of such necessity has been done using some calculations utilising the subjective scores from available image quality assessment databases. In addition, the influence of the mapping function's type has been determined as well as the increase of the PCC for the most relevant metrics and the proposed combined metrics.

The results of the conducted experiments for three most relevant image quality assessment databases are presented in Tables 1-3 and some chosen scatter plots before and after the nonlinear regression are presented in Figs. 1-6.

Table 1. Pearson correlation coefficients of the objective metrics after nonlinear mapping using various functions for the LIVE database

Mapping function	SSIM	MS-SSIM	VIF	R-SVD	RFSIM	FSIM	CQM	CISI	Hybrid FSIM	Weight FSIM
none	0.8159	0.4762	0.7327	0.4999	0.9352	0.8586	0.7214	0.9453	**0.9532**	0.8116
logistic3	0.9376	0.8929	0.7806	0.5006	0.9352	0.9539	0.7326	0.9538	**0.9559**	0.9500
logistic5	0.9446	0.6130	0.8024	0.5228	0.9352	0.9597	0.7327	0.9591	0.9572	**0.9598**
polynomial	0.9376	0.6130	0.7907	0.5089	0.9352	0.9491	0.7322	**0.9591**	0.9581	0.9320
exponential	0.9448	0.6130	0.8038	0.5102	0.9370	**0.9620**	0.7329	0.9596	0.9584	0.9500

Table 2. Pearson correlation coefficients of the objective metrics after nonlinear mapping using various functions for the TID2008 database

Mapping function	SSIM	MS-SSIM	VIF	R-SVD	RFSIM	FSIM	CQM	CISI	Hybrid FSIM	Weight FSIM
none	0.6012	0.7843	0.7777	0.4782	0.8596	0.8300	0.8600	0.8752	**0.8853**	0.8331
logistic3	0.6520	0.8390	0.8055	0.4803	0.8642	0.8710	0.8619	0.8752	**0.8853**	0.8521
logistic5	0.6542	0.8425	0.8090	0.4808	0.8645	0.8738	0.8672	0.8807	0.8873	**0.8878**
polynomial	0.6531	0.8359	0.8101	0.4808	0.8642	0.8703	0.8622	0.8790	**0.8862**	0.8855
exponential	0.6409	0.8424	0.8110	0.4866	0.8649	0.8724	0.8621	0.8807	0.8868	**0.8890**

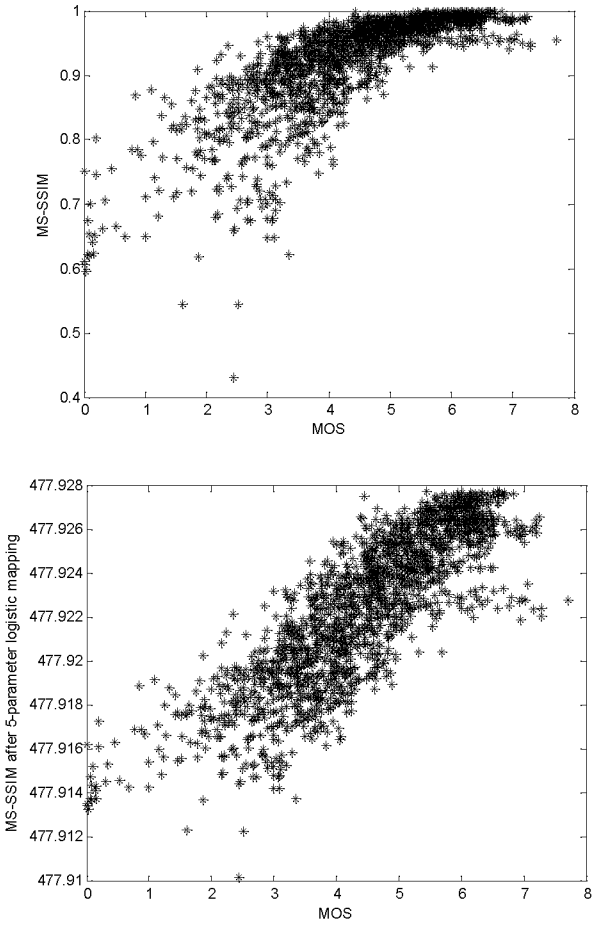

Fig. 1. Scatter plots of the MS-SSIM metric versus MOS values illustrating the changes in the linear relations between the objective and subjective scores after the nonlinear regression for TID2008 database

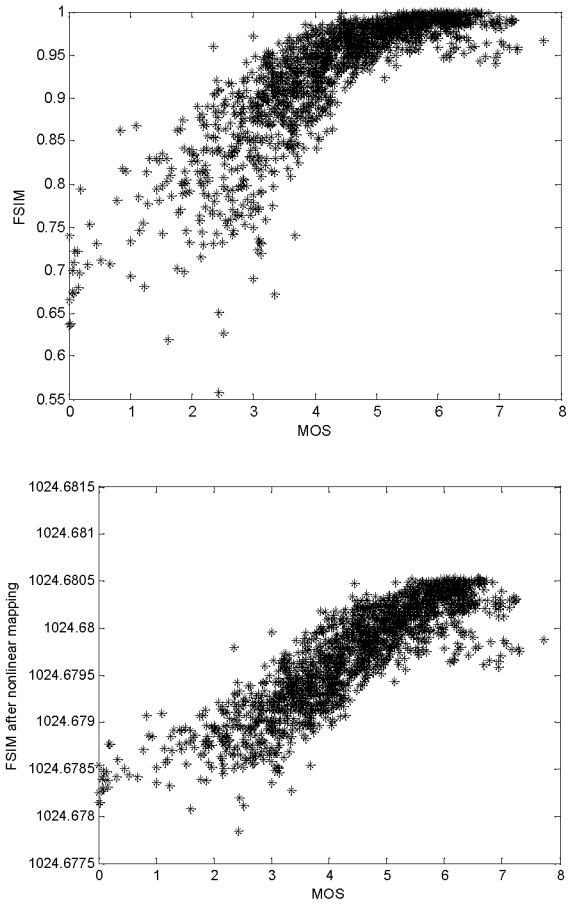

Fig. 2. Scatter plots of the FSIM metric versus MOS values illustrating the changes in the linear relations between the objective and subjective scores after the nonlinear regression for TID2008 database

Table 3. Pearson correlation coefficients of the objective metrics after nonlinear mapping using various functions for the CSIQ database

Mapping function	SSIM	MS-SSIM	VIF	R-SVD	RFSIM	FSIM	CQM	CISI	Hybrid FSIM	Weight FSIM
none	0.7654	0.7708	0.9219	0.7411	0.9130	0.8048	0.9189	**0.9346**	0.9158	0.8044
logistic3	0.8151	0.8972	0.9219	0.7437	0.9164	0.9096	0.9189	**0.9354**	0.9284	0.9220
logistic5	0.8154	0.8997	0.9278	0.7459	0.9167	0.9120	0.9208	**0.9361**	0.9284	0.9233
polynomial	0.8153	0.8872	0.9264	0.7499	0.9167	0.9028	0.9190	**0.9368**	0.9284	0.9135
exponential	0.8146	0.8998	0.9278	0.7514	0.9185	0.9104	0.9209	**0.9372**	0.9294	0.9220

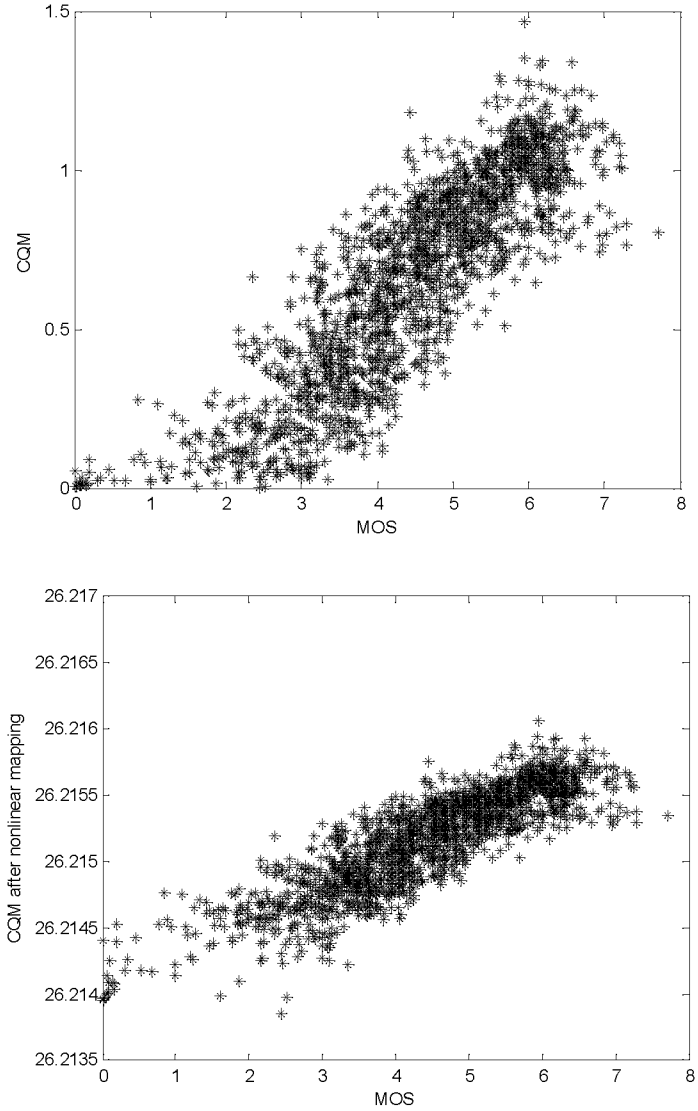

Fig. 3. Scatter plots of the CQM metric versus MOS values illustrating the changes in the linear relations between the objective and subjective scores after the nonlinear regression for TID2008 database

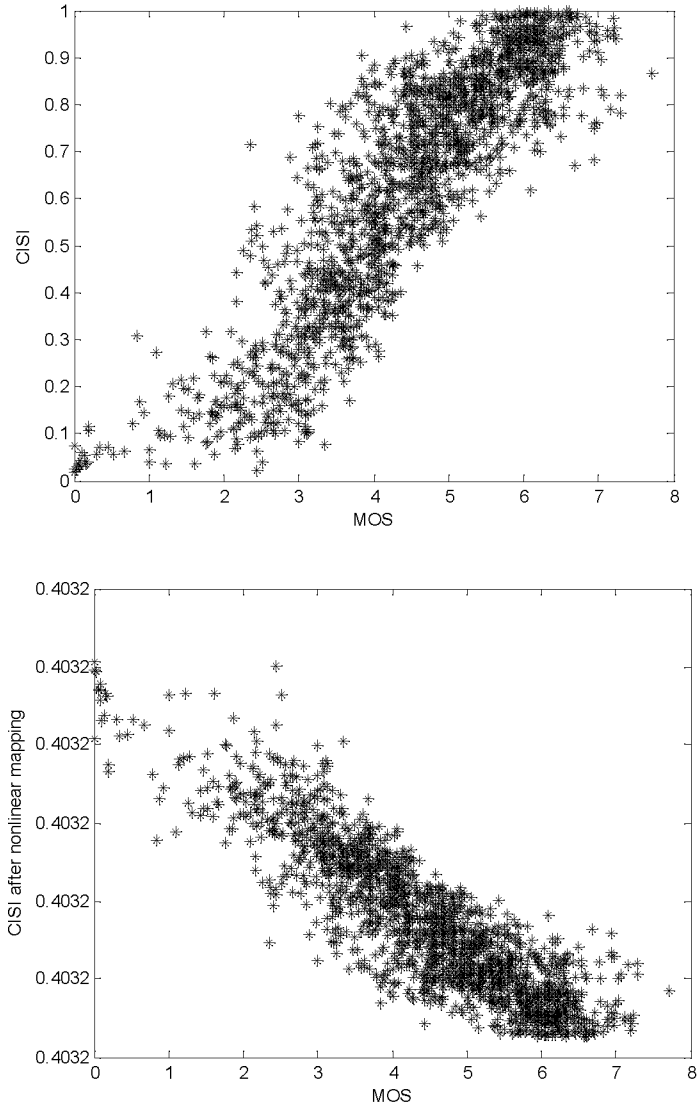

Fig. 4. Scatter plots of the CISI metric versus MOS values illustrating the changes in the linear relations between the objective and subjective scores after the nonlinear regression for TID2008 database

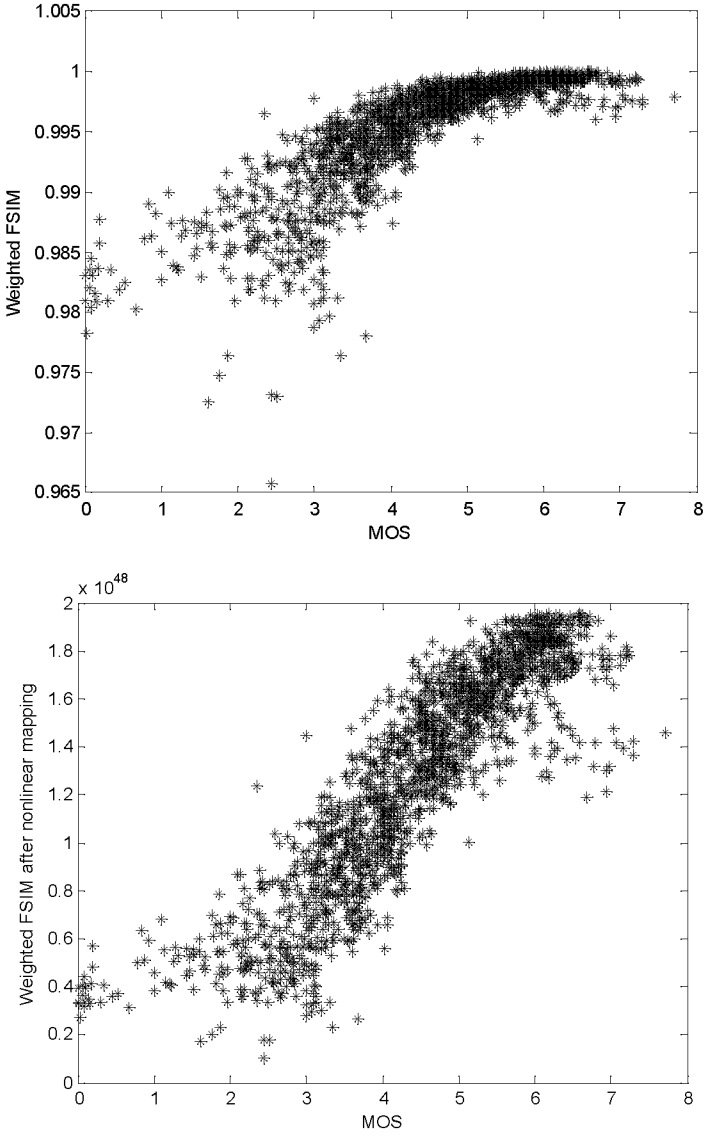

Fig. 5. Scatter plots of the Weighted FSIM metric versus MOS values illustrating the changes in the linear relations between the objective and subjective scores after the nonlinear regression for TID2008 database

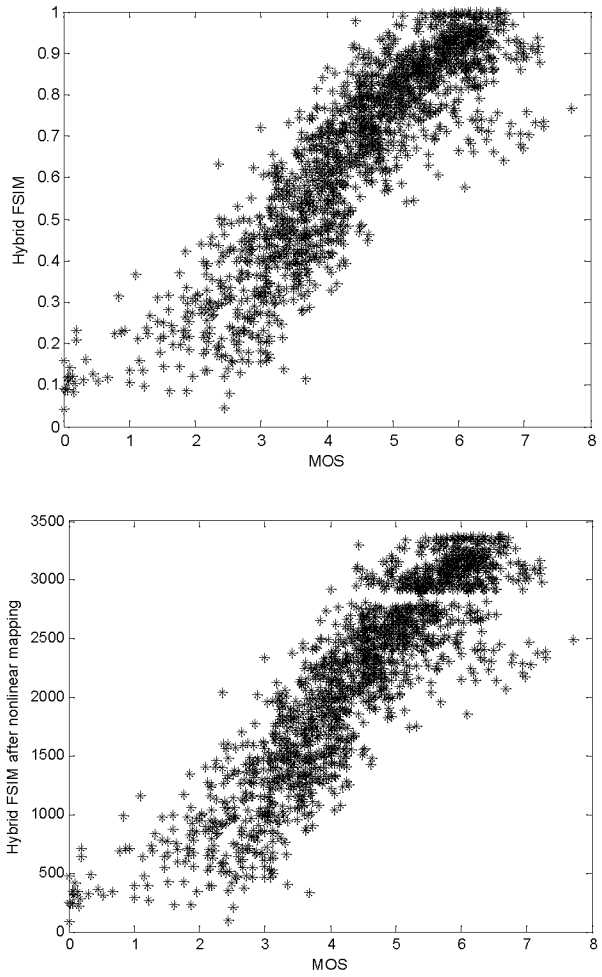

Fig. 6. Scatter plots of the Hybrid FSIM metric versus MOS values illustrating the changes in the linear relations between the objective and subjective scores after the nonlinear regression for TID2008 database

It can be noticed that the proposed combined metrics are much less sensitive to the use of the mapping function than single similarity based metrics. Analysing the results for the TID2008 database, only FSIM metric can achieve better prediction accuracy than CQM but only after nonlinear regression. Nevertheless, for CSIQ database, FSIM achieves worse results than combined metric and slightly higher correlation values can be observed for the VIF metric. Nevertheless the best results for this dataset are achieved using newly proposed Hybrid Feature Similarity and

Weighted FSIM. Slightly worse results can be obtained using Combined Image Similarity Index, which leads to the best correlation for the CSIQ dataset.

The combined metrics seem to be the representatives of the most universal approach by means of image quality prediction accuracy without the need of any additional nonlinear mapping, since the correlation coefficients obtained after the nonlinear regression are not much better, except for Weighted FSIM.

References

[1] Engelke U., Kusuma M., Zepernick H.J., Caldera M., Reduced-reference Metric Design for Objective Perceptual Quality Assessment in Wireless Imaging, Signal Processing: Image Communication, Volume 24, Number 7, pp. 525–547, 2009.

[2] Larson E., Chandler D., Most Apparent Distortion: Full-Reference Image Quality Assessment and the Role of Strategy, Journal of Electronic Imaging, Volume 19, Number 1, pp. 011006, 2010.

[3] Mansouri A., Mahmoudi-Aznaveh A., Torkamani-Azar F., Jahanshahi J., Image Quality Assessment Using the Singular Value Decomposition Theorem, Optical Review, Volume 16, Number 2, pp. 49–53, 2009.

[4] Okarma K., Combined Full-Reference Image Quality Metric Linearly Correlated with Subjective Assessment, Lecture Notes in Artificial Intelligence, Volume 6113, pp. 539–546, 2010.

[5] Okarma K., Video Quality Assessment Using the Combined Full-Reference Approach, Advances in Intelligent and Soft Computing, Volume 84, pp. 51–58, 2010.

[6] Okarma K., Colour Image Quality Assessment Using the Combined Full-Reference Approach, Advances in Intelligent and Soft Computing, Volume 95, pp. 287–296, 2011.

[7] Okarma K., Combined Image Similarity Index, Optical Review, Volume 19, Number 5, pp. 249-254, 2012.

[8] Okarma K., Weighted Feature Similarity – a Nonlinear Combination of Gradient and Phase Congruency for Full-Reference Image Quality Assessment, Advances in Intelligent Systems and Computing, Volume 184, pp. 187–194, 2013.

[9] Okarma K., Hybrid Feature Similarity Approach to Full-Reference Image Quality Assessment, Lecture Notes in Computer Science, Volume 7595, pp. 212–219, 2012.

[10] Ponomarenko N., Lukin V., Zelensky A., Egiazarian K., Carli M., Battisti F., TID2008 - a database for evaluation of full-reference visual quality assessment metrics, Advances of Modern Radioelectronics, Volume 10, pp. 30–45, 2009.

[11] Sheikh H., Bovik A., Image information and visual quality, IEEE Transactions on Image Processing, Volume 15, Number 2, pp. 430–444, 2006.

[12] Sheikh H., Wang Z., Cormack L., Bovik A., LIVE Image Quality Assessment Database Release 2, Online, 2005. http://live.ece.utexas.edu/research/quality.

[13] VQEG, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, phase I, Technical Report, 2000.

[14] VQEG, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, phase II, Technical Report, 2003.

[15] Wang Z., Bovik A., A Universal Image Quality Index, IEEE Signal Processing Letters, Volume 9, Number 3, pp. 81–84, 2002.

[16] Wang Z., Bovik A., Sheikh H., Simoncelli E., Image quality assessment: From error measurement to structural similarity, IEEE Transactions on Image Processing, Volume 13, Number 4, pp. 600–612, 2004.

[17] Wang Z., Simoncelli E., Bovik A., Multi-Scale Structural Similarity for image quality assessment, Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, California, 2003.

[18] Zhang L., Zhang L., Mou X., Zhang D., FSIM: A Feature Similarity Index for Image Quality Assessment, IEEE Transactions on Image Processing, Volume 20, Number 8, pp. 2378–2386, 2011.

[19] Zhang L., Zhang L., Mou X.: RFSIM: A Feature Based Image Quality Assessment Metric Using Riesz Transforms. Proceedings of the 17th IEEE International Conference on Image Processing (ICIP), Hong Kong, China, 2010, pp. 321–324.