Quality Assessment of Images with Multiple Distortions using Combined Metrics

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

[1]***Abstract*—In this paper the problem of quality assessment of images containing various types of distortions is concerned. Many image quality metrics proposed during last decade are quite well correlated with human perception of various kinds of distortions with the assumption that only a single type of distortions is present in the image. One of the main reasons of such approach is the lack of datasets containing subjective quality assessment results of multiply distorted images. However, after the development of LIVE Multiply Distorted Image Quality Database, a new challenge related to verification of usability of known metrics as well as the development of new ones has appeared. In this paper, the results of such verification is presented not only for some well-known metrics but also for recently proposed combined metrics together with the proposed new combined metrics optimized for multiply distorted images. The new metrics outperform previously proposed ones in the aspect of linear correlation with subjective evaluations of images containing multiple distortions.**

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

I. INTRODUCTION

Quality assessment of images subjected to various types of distortions is still one of the challenging problems in computer vision and image analysis. Depending on specific applications, also in many related areas such as robotics, bioengineering, non-destructive testing, video transmission, compression, visual inspection etc., various approaches can be applied. In some applications, where the type of possible distortions is known, some specialized metrics, sensitive to given kind of distortions, may be successfully applied even if the reference (undistorted) image is unknown. For such purposes several no-reference metrics (known also as "blind" ones) have been proposed during last decade which are useful for assessment of images contaminated by noise, blurred of lossy compressed using JPEG or JPEG2000 algorithm. Nevertheless, those methods are not universal and not necessarily very well correlated with human perception of image distortions.

In many applications, where availability of the reference image can be assumed, e.g. image compression or development of new image processing algorithms, much more universal full-reference metrics can be used. Starting from 2002 many such image quality assessment methods have been proposed by various research groups, which are more or less correlated with human perception of distortions.

Manuscript received December 7, 2013; accepted March 14, 2014.

Nevertheless, even the most recently proposed metrics are typically verified using available set of image quality assessment databases which contain numerous images subjected to various kinds of distortions and the subjective scores obtained during perceptual experiments expressed as Mean Opinion Scores (MOS) or Differential MOS (DMOS). Since such databases contain typically the reference images and a set of images with only a single type of distortion, the objective metrics are being developed and verified in the aspect of maximum correlation with available subjective scores. In the consequence, the situation when the same image is subjected to two or more types of distortions is not handled. In order to take up the challenge a new database, namely LIVE Multiply Distorted Image Quality Database [1], has been released in 2013 by a group of researchers from the Laboratory for Image and Video Engineering (LIVE), being a part of the University of Texas at Austin. Such database consists of 15 reference images and two sets of images distorted by blur followed by JPEG compression and blur followed by noise. Each distortion has been applied at four levels resulting in 16 combinations in each set and such obtained images have been assessed by nearly 20 observers. The DMOS values delivered in the database indicate that the perception of multiply distorted images differs from single distorted ones as verified in the paper [1]. For example the objective quality results obtained using BRISQUE metric [2], being one of the recently proposed "blind" metrics, using the model trained on the single distorted dataset, are too high for multiple distorted images.

II. STATE-OF-THE-ART FULL-REFERENCE METRICS

During the last several years, a number of various image and video quality assessment metrics have been proposed. Since this paper deals with full-reference metrics, due to their universality and good performance, only this kind of quality assessment methods is considered further.

Apart from "historical" metrics, such as Mean Square Error (MSE) or Peak Signal-to-Noise Ratio (PSNR), some metrics based on the structural information, better correlated with human perception, have been proposed, starting from the Universal Image Quality Index [3] and its famous extension known as Structural Similarity (SSIM) proposed about 10 years ago [4]. This metric has been also further modified (probably the most widely known extensions are Multi-Scale SSIM [5] and Gradient SSIM [6]), being also an inspiration for some other researches who have proposed some other metrics using similar structure of the calculations in some other domains. Some recent examples may be Quality Index by Local Variance (QILV) [7], Riesz-based Feature Similarity (RFSIM) [8] or Feature Similarity (FSIM) [9]. Some other types of metrics are based on the Singular Value Decomposition (one of the recent metrics of this type is known as R-SVD [10]) or information theory (IFC [11], VIF or its pixel domain version known as VIFp [12]). Some other quite popular older metrics, e.g. Visual Signal-to-Noise Ratio (VSNR), Weighted Signal-to-Noise Ratio (WSNR) or Noise Quality Measure (NQM), are included in the MATLAB based software known as MeTriX MuX package [13]. All the mentioned metrics have been used in the calculations and experiments described in this paper.

III. Motivation and Proposed Solutions

Unfortunately, currently there is no single metric which would significantly outperform many others, so a natural consequence of this fact is the new direction of research related to the fusion of different metrics into some combined ones, even at the cost of higher computational complexity.

Another disadvantage of most of known objective metrics is their nonlinear relationship with perception of various distortions by human observers. A partial solution of this problem, as suggested by the Video Quality Experts Group (VQEG), is the nonlinear mapping of scores using the logistic or exponential functions in order to linearize the relationship between the metric and MOS or DMOS values. Unfortunately, the coefficients of the mapping functions are different for various datasets, so the universality of such approach is rather low.

The first successful attempt to the application of combined metrics is the idea presented in the paper [14], where the proposed combination of three metrics (MS-SSIM, VIF and R-SVD) leads to the Pearson's linear correlation coefficient (PCC) equal to 0.86 for the most relevant Tampere Image Database (TID2008) containing 1700 distorted images with 17 types of distortions assessed by 838 observers [15]. The PCC values for the single metrics are much lower (max. 0.784 for MS-SSIM without nonlinear mapping) for the same dataset.

Further modifications based on applying some newer metrics and combination of them presented in some recent papers [16], [17] have verified the usefulness of such approach and lead to even better results. A recently proposed [18] combination of MS-SSIM, VIF, RFSIM and modified FSIMc metric, named the Extended Hybrid Image Similarity (EHIS) ensured the PCC = 0.9105 for TID2008 dataset.

The validity of this approach has been also justified by some other researchers, e.g. in one of the most recent articles [19], where quite deep analysis of various combinations of metrics has been presented. Nevertheless, the fused methods presented in this paper are less universal as they assume the context classification based on the knowledge of the distortion type and the training process is conducted separately for each type of distortions. It is also worth to notice that additional logistic mapping has been also applied to fit the objective scores to the DMOS values. Although its impact is high mainly at the extremes of the test range, related to very high or extremely low quality images, changes of the PCC values may also be noticed.

Based on the idea of the combined metrics and the specific character of the multiply distorted image, the verification of the validity of the combined metrics for such images becomes a challenge taken in the research presented in this paper together with the results of the optimization of the combined metrics dedicated for multiply distorted images.

IV. Experiments and Obtained Results

Considering the structure of the LIVE Multiply Distorted Image Quality Database built from two image sets, the analysis of the correlation of chosen metrics with subjective scores, has been conducted in two ways. First, the calculations have been done using MATLAB environment with necessary toolboxes for all 450 distorted images (225 in each set) and the correlation coefficients have been computed for them without any nonlinear fitting. Then, due to the presence of single distorted images (6 of 15 distorted images for each reference one) in the database, the set of analysed images has been limited to multiply distorted images (remaining 270 images) and the same computations have been conducted. The PCC values obtained in both cases are presented in Table I.

TABLE I. PEARSON'S LINEAR CORRELATION COEFFICIENTS WITH DMOS VALUES OBTAINED FOR VARIOUS FULL-REFERENCE METRICS.

Metric	PCC for all 450 images	PCC for the set of 270 multiply distorted images
SSIM	0.6683	0.4666
MS-SSIM	0.8062	0.6429
GSSIM	0.8113	0.4753
QILV	0.6222	0.4821
IFC	0.8418	0.7769
VIF	0.7586	0.7163
VIFp	0.7746	0.6614
R-SVD	0.7678	0.4118
NQM	**0.8985**	**0.8083**
VSNR	0.8335	0.5418
WSNR	0.8147	0.5213
RFSIM	0.8705	0.6578
FSIM	0.8185	0.7076
WFSIM	0.7772	0.6833
CQM [14]	0.8804	0.7112
CISI [16]	0.8967	0.7567
EHIS [18]	0.7493	0.6094

Analysing the presented results, it can be easily noticed that much better results can be achieved for the whole dataset due to relatively high correlation of all metrics with subjective scores of single distorted images (blurred and JPEG compressed), especially for such typical distortions. Surprisingly, in both cases the best results (even better than for the combined metrics) are obtained for the Noise Quality Measure. Especially high correlation with DMOS values in comparison to the other metrics, can be observed for the set of multiply distorted images.

Considering the results obtained for the single metrics, the optimization of the combined metrics constructed from two different ones has been conducted. A general form of the two-component combined metric is expressed as

$$CM2 = \left(Metric1\right)^{a} \times \left(Metric2\right)^{b}, \quad (1)$$

where a and b are the weighting exponents being optimized towards maximum linear correlation of the combined metric with DMOS values. Obviously, such combined metrics can also be built from three or more different metrics.

As the next part of the experiments the optimization of all pairs of the metrics has been conducted for the set of multiply distorted images. Similar calculations have been performed for the whole dataset leading to the same conclusions. The PCC values with DMOS values obtained for the two-component combined metrics with optimized exponents are presented in Table II.

Analysing the results a great importance of the choice of the metrics can be easily noticed. Since the highest correlation with subjective scores has been obtained for the combination of NQM and IFC metrics, they have been chosen as the basic ones for further extensions. Adding the third metric to the combined one, the results presented in the left part of the Table III have been achieved with the best result for the combination with the VSNR. The obtained metric can be expressed as

$$CM3 = IFC^{0.34} \times NQM^{2.4} \times VSNR^{-0.3}, \quad (2)$$

where the four-component metric obtained after optimization (assuming the presence of IFC, NQM and VSNR components) is

$$CM4 = IFC^{0.2} \times NQM^{2.9} \times VSNR^{-0.54} \times VIF^{0.5}. \quad (3)$$

The linearity of the relationship between DMOS and the proposed metrics is illustrated additionally by the scatter plots presented in Fig. 1 and Fig. 2, where different values of DMOS values can be noticed for images subjected to single and multiple distortions. Most of multiply distorted images have the DMOS values higher than 30 or even 40 whereas over the half of single distorted images have lower DMOS values.

The main advantage of the highly linear relation between those DMOS values and the proposed metric is the possibility of an accurate prediction of the perceived image quality by calculation of the values of the combined metrics.

The detailed results of the linear correlation coefficients obtained without any nonlinear mapping for four-component metrics are presented in the right part of Table III.

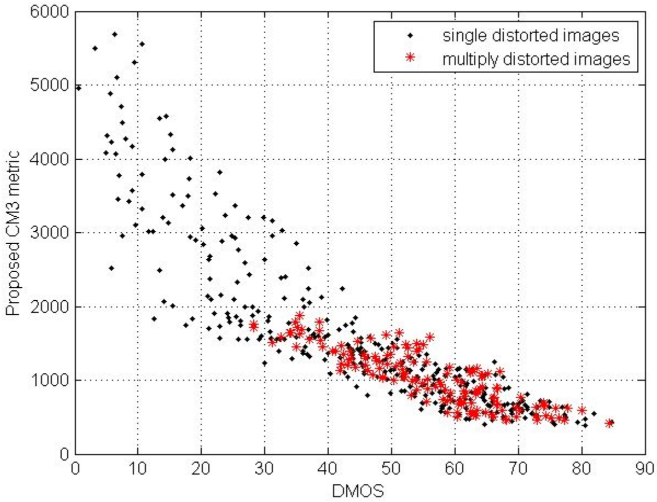

Fig. 1. Scatter plot of the CM3 metric versus DMOS values for the LIVE Multiply Distorted Image Quality Database.

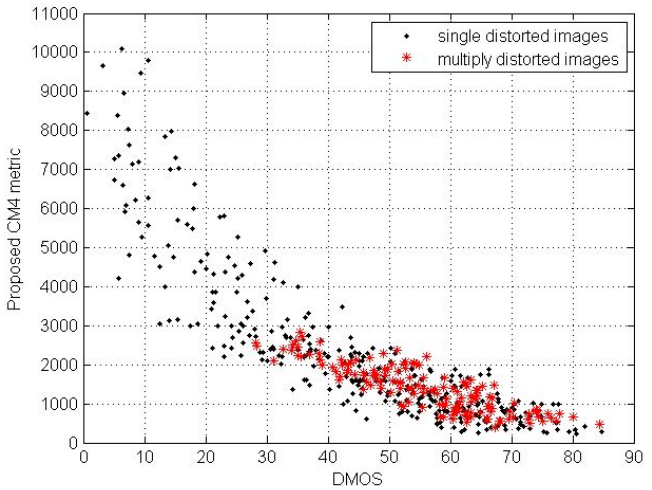

Fig. 2. Scatter plot of the CM4 metric versus DMOS values for the LIVE Multiply Distorted Image Quality Database.

TABLE II. PEARSON'S LINEAR CORRELATION COEFFICIENTS WITH DMOS VALUES OBTAINED FOR TWO-COMPONENT COMBINED METRICS FOR ALL IMAGES (UPPER RIGHT VALUES) AND FOR 270 MULTIPLY DISTORTED IMAGES (LOWER LEFT VALUES).

450 images / **270 images**	**SSIM**	**MS-SSIM**	**GSSIM**	**QILV**	**IFC**	**VIF**	**VIFp**	**R-SVD**	**NQM**	**VSNR**	**WSNR**	**RFSIM**	**FSIM**
SSIM	---	0.8892	0.8152	0.8213	0.8899	0.7770	0.7880	0.8155	0.9060	0.8446	0.8400	0.8746	0.8971
MS-SSIM	0.7227	---	0.8757	0.8748	0.9003	0.8766	0.8768	0.8893	0.9072	0.8772	0.8823	0.8844	0.8970
GSSIM	0.4941	0.7200	---	0.8404	0.8896	0.8195	0.8154	0.8382	0.9080	0.8501	0.8487	0.8767	0.8961
QILV	0.5421	0.6910	0.5421	---	0.8981	0.8519	0.8383	0.8422	0.9057	0.8376	0.8238	0.8752	0.9005
IFC	0.7817	0.7869	0.7813	0.7972	---	0.8896	0.8896	0.8966	**0.9229**	0.9003	0.9013	0.9102	0.9094
VIF	0.7436	0.7382	0.7350	0.7865	0.7972	---	0.7847	0.8258	0.9120	0.8593	0.8569	0.8860	0.9069
VIFp	0.7191	0.6910	0.6871	0.7831	0.7865	0.7324	---	0.8385	0.9089	0.8542	0.8569	0.8802	0.8997
R-SVD	0.6051	0.7163	0.6016	0.8454	0.7831	0.7396	0.7040	---	0.9162	0.8757	0.8421	0.8871	0.8996
NQM	0.8138	0.8142	0.8150	0.7976	**0.8454**	0.8366	0.8255	0.8200	---	0.9029	0.9055	0.9049	0.9073
VSNR	0.5707	0.6918	0.5747	0.8018	0.7976	0.7298	0.6735	0.6469	0.8248	---	0.8509	0.8725	0.8932
WSNR	0.5458	0.6900	0.5527	0.8110	0.8018	0.7301	0.6685	0.5814	0.8133	0.5889	---	0.8728	0.8929
RFSIM	0.6582	0.6918	0.6602	0.8033	0.8110	0.7403	0.7010	0.6933	0.8140	0.6718	0.6582	---	0.8934
FSIM	0.7445	0.7445	0.7444	0.7972	0.8033	0.7702	0.7486	0.7494	0.8163	0.7509	0.7650	0.7469	---

TABLE III. PEARSON'S LINEAR CORRELATION COEFFICIENTS WITH DMOS VALUES CALCULATED FOR 270 MULTIPLY DISTORTED IMAGES OBTAINED FOR VARIOUS THREE-COMPONENT METRICS.

Three-component metrics		Four-component metrics	
IFC+NQM+	**PCC**	**IFC+NQM+VSNR+**	**PCC**
SSIM	0.8468	SSIM	0.8512
MS-SSIM	0.8480	MS-SSIM	0.8512
GSSIM	0.8474	GSSIM	0.8512
QILV	0.8496	QILV	0.8522
VIF	0.8493	VIF	**0.8596**
VIFp	0.8457	VIFp	0.8559
R-SVD	0.8463	R-SVD	0.8515
RFSIM	0.8458	RFSIM	0.8562
FSIM	0.8456	FSIM	0.8536
WSNR	0.8457	WSNR	0.8545
VSNR	**0.8512**		

V. CONCLUSIONS

The results presented in the paper have proven that an automatic assessment of multiply distorted images is still one of the challenges of computer vision and image analysis. Since the accuracy of the quality prediction is more important than the prediction monotonicity, the optimization of objective metrics is based on the Pearson's linear correlation coefficient (directly related to the prediction accuracy) and rank-order correlation (Spearman and Kendall ones) are considered as supplementary ones.

As we can observe on the scatter plots and the PCC values presented in Tables I-III, the subjective perception of multiply distorted images differs significantly from the assessment of images contaminated by a single type of distortions. It is also justified by the fact that the highest correlation with DMOS values can be obtained using the nonlinear combination of different metrics than previously proposed for single distorted images. However, the application of the combined metrics for multiply distorted images seems to be an interesting idea as it leads to significant increase of the quality prediction accuracy in comparison to the single full-reference metrics.

Although the results presented in the paper can be treated a bit preliminary due to the relatively small number of images and distortion types present in the only available dataset containing the subjective scores of images subjected to multiple distortions, they can be a good starting point for further extensive research.

Nevertheless, it is worth to notice that the development of any objective metric highly correlated with human perception of multiple distortions will be still limited by availability of appropriate databases.

REFERENCES

[1] D. Jayaraman, A. Mittal, A. K. Moorthy, A. C. Bovik, "Objective image quality assessment of multiply distorted images", in *Conf. Rec. 46th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, California, 2012, pp. 1693–1697. [Online]. Available: http://dx.doi.org/10.1109/ACSSC.2012.6489321

[2] A. Mittal, A. K. Moorthy, A. C. Bovik, "No-reference image quality assessment in the spatial domain", *IEEE Trans. on Image Processing*, vol. 21, no. 12, pp. 4659–4708, 2012. [Online]. Available: http://dx.doi.org/10.1109/TIP.2012.2214050

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

[4] 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, 2004. [Online]. Available: http://dx.doi.org/10.1109/TIP.2003.819861

[5] 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. [Online]. Available: http://dx.doi.org/10.1109/ACSSC.2003.1292216

[6] G.-H. Chen, C.-L. Yang, S.-L. Xie, "Gradient-based structural similarity for image quality assessment", in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, Altanta, Georgia, 2006, pp. 2929–2932. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2006.313132

[7] S. Aja-Fernandez, R. San-Jose-Estepar, C. Alberola-Lopez, C. Westin, "Image quality assessment based on local variance", in *Proc. 28th Annu. IEEE Int. Conf. Engineering in Medicine and Biology Society (EMBS 2006)*, New York, 2006, pp. 4815–4818. [Online]. Available: http://dx.doi.org/10.1109/IEMBS.2006.259516

[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. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2010.5649275

[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, 2011. [Online]. Available: http://dx.doi.org/10.1109/TIP.2011.2109730

[10] A. Mansouri, A. Mahmoudi-Aznaveh, F. Torkamani-Azar, J. A. Jahanshahi, "Image quality assessment using the singular value decomposition theorem", *Opt. Rev.*, vol. 16, no. 2, pp. 49–53, 2009. [Online]. Available: http://dx.doi.org/10.1007/s10043-009-0010-y

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

[12] H. Sheikh, A. Bovik, "Image information and visual quality", *IEEE Trans. on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006. [Online]. Available: http://dx.doi.org/10.1109/TIP.2005.859378

[13] M. Gaubatz, "MeTriX MuX visual quality assessment package", [Online]. Available: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[14] K. Okarma, "Combined full-reference image quality metric linearly correlated with subjective assessment", in *Proc. 10th Int. Conf. Artificial Intelligence and Soft Computing, (ICAISC 2010)*, Zakopane, Poland, 2010, pp. 539–546. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13208-7_67

[15] 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, p. 30–45, 2009.

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

[17] K. Okarma, "Hybrid feature similarity approach to full-reference image quality assessment", in *Proc. Int. Conf. Computer Vision and Graphics (ICCVG 2012)*, Warsaw, Poland, 2012, pp. 212–219. [Online]. Available: http://dx.doi.org/10.1007/ 978-3-642-33564-8_26

[18] K. Okarma, "Extended hybrid image similarity – combined full-reference image quality metric linearly correlated with subjective scores", *Elektronika ir Elektrotechnika*, vol. 19, no. 10, pp. 129–132, 2013. [Online]. Available: http://dx.doi.org/10.5755/j01.eee.19.10.5908

[19] T-J. Liu, W. Lin, C-C. J. Kuo, "Image quality assessment using multi-method fusion", *IEEE Trans. Image Processing*, vol. 22, no. 5, pp. 1793–1807, 2013. [Online]. Available: http://dx.doi.org/ 10.1109/TIP.2012.2236343