https://doi.org/10.2352/ISSN.2470-1173.2018.13.IPAS-260

Robust linearized combined metrics of image visual quality

Oleg Ieremeiev[a], Vladimir Lukin[a], Nikolay Ponomarenko[a], Karen Egiazarian[b]
[a] National Aerospace University, 61070, Kharkov, Ukraine;
[b] Tampere University of Technology, FIN 33101, Tampere, Finland

Abstract

Existing full-reference metrics still do not provide a desirable degree of adequacy to a human visual perception, for evaluation of images with different types and levels of distortions. One reason for this is that it is difficult to incorporate the peculiarities of human visual system in the metrics design. In this paper, a robust approach to full-reference metrics' design is proposed, based on a combination of several existing full-reference metrics. A preliminary linearization (fitting) of the dependence of MOS with respect to the components metrics is performed in order to compensate shortcomings of each component. The proposed method is tested on several known databases, and demonstrate better performance than existing metrics.

Keywords: image visual quality assessment, full-reference metrics, combined metrics, robust metrics

Introduction

A volume and resolutions of acquired images, which need to be transmitted, processed, stored and disseminated, are continuously increasing. Since most of images are subject to a visual inspection and analysis, the methods of objective assessment of original image quality and efficiency of image processing application that take into account peculiarities of human vision system (HVS) become actual [1-3]. This leads to a rapid development of so-called visual quality metrics including so-called full-reference (FR) quality metrics used for the verification of the results of various image processing tasks, such as denoising, deblurring, lossy compression, super-resolution, etc. Based on such an assessment, it is possible to optimize (adjust) parameters of image processing algorithms [2-4], etc.

The main problem with the most of existing metrics is that they use simplified mathematical models of human visual perception. Due to this, even the best performing metrics on the largest databases of test images (for example, TID2013) show low values of the Spearman rank order correlation coefficient (SROCC, determined between mean opinion score (MOS) and a metric) reaching only the level of 0.83-0.85 [3]. One way to improve metrics' performance is to combine them. Several ways to combine visual quality metrics have been proposed in literature [5-10]. This can be done using simple functions, with elementary metrics used as the arguments [5]. Other ways to combine is to employ some decision rules [9], clustering, neural networks or other learning techniques [7, 8, 10]. On one hand, more complex ways of combining elementary metrics lead to better results (i.e. a larger correlation of a metric with the mean opinion score for certain databases of distorted images) [10]. On the other hand, in practice, one needs a simple but yet universal metric. This restricts the number of elementary visual quality metrics used jointly in a combined metric.

Existing visual quality metrics have different nonlinear dependencies with MOS and different ranges of metric values. There are metrics (e.g. PSNR, expressed in dBs) that have unlimited range [6], whilst there are other metrics that have limited ranges (e.g., from 0 to 1), and, moreover, metric values mostly concentrate only in a part of this range [11]. Besides, MOS can vary in different limits [1, 11]. To get around aforementioned inconsistency, different approaches have been used. Fitting with a further linearization is one of them [7].

The combined metrics have to be optimized or trained, and it is important to choose a proper set of test images for it. Since the main goal of this study is to create a "universal" metric which is robust to some types of image distortions and their combinations, the combined metric should be optimized for the large databases containing various types of distortions, such as blur, noise (Poisson, spatially correlated, impulse and others), artifacts caused by filtering or lossy compression, etc. Thus, a metric design and optimization requires a large image database. An example of such a database is TID2013 [3]. Note, that the metrics which are optimized for one database should also perform well for other databases.

Thus, the goal of this paper is to design a simple and universal framework of combined visual quality metrics that have quasi-linear dependence with MOS and outperform existed metrics on different databases.

Proposed approach to a combined metric design

As it has been mentioned above, image quality metrics may have different ranges of variations of metric values. In this regard, the MOS to metric fitting with a linearization offers opportunities of providing possible transformation of a metric value into the corresponding MOS value or a value proportional to MOS. This operation is simple and can be done in advance for any elementary metric obtaining a proper dependence (e.g., power function) for each component – elementary visual quality metric. After such a transformation, one, in fact, has the number of MOS estimates as the number of component metrics considered. These MOS estimates can be further processed linearly or nonlinearly. In this paper, we follow the latter approach. Since we consider a limited number of elementary metrics (up to five) to ensure a simple structure of a combined metric, we have a limited number of variants for robust processing of MOS estimates, based on, e.g., sample median or alpha-trimmed mean.

The main idea of the proposed approach is in the following. Recall that there are no elementary visual quality metrics that perform equally well for different types of degradations in images. One metric can over-estimate quality of images for one type of degradation whilst it under-estimates quality for images with other type or types of distortions. In other words, a given visual quality metric can produce obviously or sufficiently erroneous estimate of MOS. Then, if the different metrics have different advantages and shortcoming (week points, "unfavorable" types of distortions), then their joint processing in the form of linearized MOS using

robust estimation shall lead to a decrease of erroneous estimate influence. This property diminishes the drawbacks of elementary metrics (inaccurate assessment of images for a particular type or types of distortions).

Below, we consider the robust linearized combined metrics designed for the following three configurations: median of MOS for three component metrics, median of MOS for five component metrics, alpha-trimmed mean of five MOS estimates. Optimization presumes finding the best set of component metrics among available ones. Before describing an optimization procedure and its results, let us consider more practical but still an important task of linearization.

Linearization of the elementary metrics

In general, there are number of ways to linearize dependencies presented in the form of scatter-plots (MOS vs a metric in our case). For example, it is possible to apply exponential, logarithmic, power functions as well as polynomials of different order. Let us give more details and requirements to fitting and linearization, keeping in mind that they can influence further stages of joint processing. Recall that a linearization (and fitting) also depends on amount of data and a way scatter-plot is obtained. Accuracy of fitting can be described by different parameters and this should be taken into account as well.

To have a variety of distortion types and many points of our scatter-plots, we have exploited the database TID2013 [3] which is currently the largest open database according to the number of distorted images (3000), analyzed types of distortions (24) and volunteers participating in the experiments (about 1000) used to obtain MOS.

To explain what is expected from the fitting and what is undesirable, let us consider several examples presented for different metrics in Fig. 1. Recall here that MOS for images in the database TID2013 potentially varies from 0 to 9 and, in fact, limits of variations are from almost zero (very bad quality) to slightly larger than 7 (perfect quality without visible distortions). Horizontal axis in all scatter-plots corresponds to a considered metric whilst vertical axis corresponds to MOS. Note that we expect that dependence of MOS on a metric is monotonous, i.e. a larger value of a metric (desirably) relates to a larger MOS associated with better visual quality.

Our examples in Fig. 1 relate to two groups of visual quality metrics that are most popular nowadays. They are based on (are extensions of) either conventional PSNR or SSIM [12]. Among these extensions, we have considered those metrics that have the largest values of Spearman rank order correlation coefficient (SROCC) [3] with MOS. PSNR is calculated as:

$$\mathrm{PSNR} = 10 \log_{10}(255^2/\mathrm{MSE}) \ , \qquad (1)$$

where MSE is the mean square error for pixel-wise comparison of distorted and reference images. Modifications mainly relate to calculation of modifications of MSE [6, 13]. All these metrics are expressed in dB with larger values corresponding to better visual quality.

SSIM-based metrics “analyze” similarity between two images (distorted and reference ones) using correlation analysis. Their values vary from 0 to 1, but, in fact, are mainly concentrated in the limits from 0.8 to 1.0 (see data in Fig. 1).

First, it is seen that the fitted curve is not always monotonous (we have used Curve Fitting Toolbox that provides wide opportunities of employing different functions for this purpose). The examples are dependencies of MOS on SSIM [12] for the functions Exp2 and Fourier3, dependence of MOS on PSNRHMAm [13] (which is modification of PSNR-HMA [6]) for the functions Gauss1 and Fourier3, for MOS on FSIMc [4] for the function Poly4. There can be also the case when the fitted curve is monotonous but it does not suit the considered task – this happens for the dependence of MOS on PSNRHMAm for the function Exp1 (the MOS values can be larger than 9). Meanwhile, there are also examples (Fig. 2) of intuitively good approximations as MOS on PSNR for the function Power1.

Second, it becomes clear that one type of approximating functions can be, in general, better than others. In particular, polynomials of high order (e.g., >6) can overfit the data. Besides, it is worth applying some quantitative measures of fitting quality. This can be, for example, conventional Pearson correlation (PC) or root mean square error (RMSE) (recall that SROCC does not change in cases of monotonous fitting).

The fitting results for four aforementioned metrics are presented in Table 1. As it is seen, the best fitting for the metric PSNRc according to both quantitative criteria (maximal PC and minimal RMSE) has been provided by the fifth order polynomial, although the difference in fitting quality between all considered variants is not large. The best fitting for the metric PSNRHMAm is observed for polynomials of the third and fourth order according to RMSE and Fourier functions according to PC.

SSIM is the best approximated by the fifth order polynomial. The same relates to FSIMc.

According to both accuracy measures, the best fit takes place for the metric PSNRHMAm. Note that in the case of the best fit the values of PC are approximately the same as the corresponding values of SROCC presented in the last row.

Fig. 2 also gives an example of three types of fitting functions where the variant Power2 is obviously the best one according to the quantitative criteria and visual analysis. Keeping in mind this property and recommendations given in [8], we have obtained fitting results for fitting function Power2.

Robust linearized metrics

Suppose now that one has several estimates of MOS obtained by linearization of different metrics for the database TID2013. The chosen metrics are those analyzed in [3] that are among the best elementary metrics. The obtained data are presented in Table 2. PC and RMSE values are given to characterize the quality of fitting. Besides, all three parameters of the fitting function are given in three rightmost columns. SROCC values are presented as well.

As it is seen, the largest PC values are attained for the metrics FSIMc, SFF [14], and PSNRHMSm that also have the largest SROCC and the smallest RMSE values. While selecting the metrics for joint processing (to design a combined metric), we should use good but different (complementing each other) metrics.

If one uses three metrics jointly, there are numerous ways to choose them from the set of elementary metrics given in Table 2. Table 3 presents five best results (totally 1771 combinations have been studied) obtained for three visual quality metrics for which final MOS has been obtained as the median of three MOS estimates from three different elementary metrics after linearization. Optimization (selection of the best sets) has been carried out for TID2013 database using two criteria – either SROCC or standard Pearson correlation. According to these criteria, the obtained results are similar.

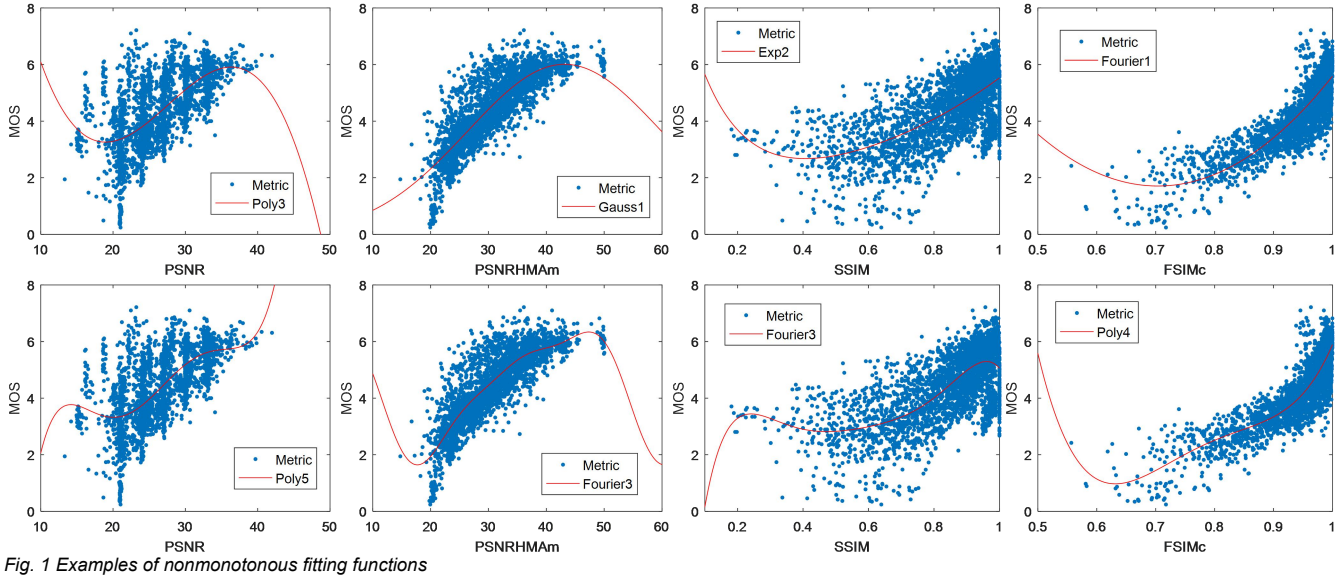

Fig. 1 Examples of nonmonotonous fitting functions

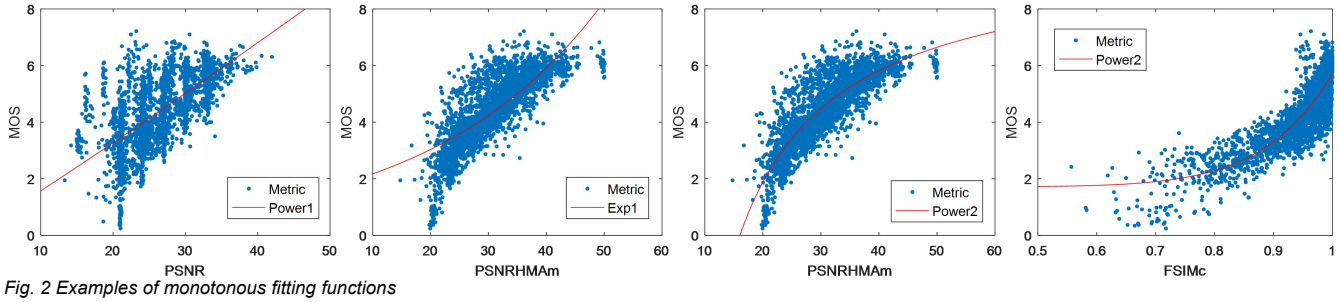

Fig. 2 Examples of monotonous fitting functions

Table 1. Fitting functions and accuracy of fitting

Notation	Function	PSNR		PSNRHMAm		SSIM		FSIMc	
		RMSE	PC	RMSE	PC	RMSE	PC	RMSE	PC
Exp1	**f(x) = a*exp(b*x)**	**0.930**	**0.662**	**0.750**	**0.798**	**0.923**	**0.668**	**0.615**	**0.869**
Exp2	**f(x) = a*exp(b*x) + c*exp(d*x)**	**0.930**	**0.662**	**0.817**	**0.811**	**0.913**	**0.677**	**0.596**	**0.877**
Fourier1	f(x) = a0 + a1*cos(x*w) + b1*sin(x*w)	0.911	0.679	0.636	0.859	0.902	0.686	0.617	0.868
Fourier2	f(x) = a0 + a1*cos(x*w) + b1*sin(x*w) + a2*cos(2*x*w) + b2*sin(2*x*w)	0.908	0.682	0.635	0.859	0.896	0.692	0.595	0.877
Fourier3	f(x) = a0 + a1*cos(x*w) + b1*sin(x*w) + a2*cos(2*x*w) + b2*sin(2*x*w) + a3*cos(3*x*w) + b3*sin(3*x*w)	0.905	0.685	0.634	0.860	0.892	0.696	0.595	0.877
Gauss1	f(x) = a1*exp(-((x-b1)/c1)^2)	0.928	0.663	0.642	0.856	0.923	0.668	-	-
Poly1	**f(x) = p1*x + p2**	**0.904**	**0.660**	**0.614**	**0.832**	**0.958**	**0.652**	**0.687**	**0.832**
Poly2	f(x) = p1*x^2 + p2*x + p3	0.903	0.663	0.556	0.859	0.918	0.676	0.603	0.867
Poly3	f(x) = p1*x^3 + p2*x^2 + p3*x + p4	0.879	0.678	0.553	0.859	0.917	0.683	0.564	0.869
Poly4	f(x) = p1*x^4 + p2*x^3 + p3*x^2 + p4*x + p5	0.878	0.68	0.548	0.858	0.907	0.692	0.560	0.877
Poly5	f(x) = p1*x^5 + p2*x^4 + p3*x^3 + p4*x^2 + p5*x + p6	0.875	0.681	0.548	0.858	0.903	0.694	0.560	0.877
Power1	**f(x) = a*x^b**	**0.931**	**0.661**	**0.695**	**0.829**	**0.939**	**0.654**	**0.625**	**0.864**
Power2	**f(x) = a*x^b+c**	**0.928**	**0.663**	**0.641**	**0.856**	**0.917**	**0.673**	**0.603**	**0.874**
	SROCC		0.687		0.854		0.637		0.851

Table 2. Characteristics of metrics after linearization

Metrics	SROCC	Power2		f(x) = a*x^b+c		
		PC	RMSE	a	b	c
PSNRc	0.687	0.663	0.928	0.006518	1.817	1.838
MSSIM	0.787	0.834	0.685	3.774	8.777	1.823
SSIM	0.637	0.673	0.917	3.175	2.738	2.358
VSNR	0.681	0.661	0.930	-44.9	-1.051	6.09
VIF	0.677	0.767	0.796	6.288	0.3041	-0.785
VIFP	0.608	0.712	0.871	-7.148	-0.1762	12.62
NQM	0.635	0.604	0.989	-14	-0.208	11.82
WSNR	0.579	0.521	1.059	0.6048	0.5245	0.642
IFC	0.540	0.632	0.961	-2.426	-0.4234	5.788
IWSSIM	0.777	0.829	0.693	3.681	6.774	1.842
CWSSIM	0.562	0.563	1.025	1.679	2408	3.588
DCTUNE	0.620	0.610	0.982	5.535	-0.3485	2.284
MAD_INDEX	0.781	0.819	0.711	-0.004784	1.369	5.637
PSNRHVS	0.653	0.590	1.001	-56.14	-0.9029	7.272
PSNRHVSM	0.624	0.590	1.002	-22.85	-0.4282	9.741
PSNRHAc	0.819	0.833	0.686	-362.5	-1.342	8.244
PSNRHMAc	0.813	0.829	0.693	-210	-1.133	8.601
FSIM	0.801	0.857	0.640	4.042	8.627	1.542
FSIMc	0.851	0.874	0.603	3.999	8.713	1.719
SRSIM	0.807	0.865	0.623	4.115	14.65	1.45
SFF	0.851	0.868	0.616	4.089	16.76	1.675
PSNRHMAm	0.854	0.856	0.641	-131.4	-0.9193	10.25

Table 3 – Median of three estimates of MOS

##	Metrics	SROCC	PC
1	**FSIMc, SFF, PSNRHMAm**	**0.8794**	**0.8964**
2	SRSIM, SFF, PSNRHMAm	0.8775	0.8953
3	MAD_INDEX, SFF, PSNRHMAm	0.8765	0.8950
4	FSIM, SFF, PSNRHMAm	0.8757	0.8935
5	DCTUNE, FSIMc, SFF	0.8742	0.8915

As one can see, the use of the median for MOS of three elementary metrics leads to SROCC and Pearson correlation coefficient about 0.89 for several combinations. There are several sets producing similar results. Thus, we have the benefit of about 0.02 for both SROCC and PC.

The use of five component metrics (totally more than 33000 combinations have been studied) allows further increasing of SROCC and Pearson correlation to almost 0.9 (see data in Table 4).

These results can be still improved a little if α-trimmed estimate (Table 5) is applied (the largest and the smallest MOS estimates are rejected and three remained ones are averaged).

Therefore, sufficient improvement compared to elementary metrics (and compared to the best metric among them) has been attained.

Table 4 – Median of five estimates of MOS

##	Metrics	SROCC	PC
1	**VIFP, DCTUNE, FSIMc, SFF, PSNRHMAm**	**0.8847**	**0.9022**
2	SSIM, DCTUNE, FSIMc, SFF, PSNRHMAm	0.8844	0.8989
3	IFC, DCTUNE, FSIMc, SFF, PSNRHMAm	0.8839	0.9016
4	UQI, DCTUNE, FSIMc, SFF, PSNRHMAm	0.8838	0.9003
5	DCTUNE, VIF, FSIMc, SFF, PSNRHMAm	0.8831	0.9018

Table 5 - α-trimmed mean

##	Metrics	SROCC	PC
1	**IFC, DCTUNE, FSIMc, SFF, PSNRHMAm**	**0.8871**	**0.9053**
2	SSIM, DCTUNE, FSIMc, SFF, PSNRHMAm	0.8871	0.9001
3	VIFP, DCTUNE, FSIMc, SFF, PSNRHMAm	0.8862	0.9045
4	UQI, DCTUNE, FSIMc, SFF, PSNRHMAm	0.8851	0.9013
5	DCTUNE, VIF, FSIMc, SFF, PSNRHMAm	0.8847	0.9044

A question is what has one to pay for this improvement. Obviously, it becomes necessary to determine three or five elementary metrics instead of one, then to carry out determination of three or five MOS estimates using linearization and to apply the considered robust estimator. The latter two operations are simple and do not require essential efforts. Determination of elementary metrics can be done in parallel and then time for their calculation is approximately the same as for the computationally most complex metrics.

It is worth stressing that the best combinations in Tables 3, 4, and 5 are based on elementary metrics that incorporate different principles. For example, the best combination in Table 3 includes FSIMc, SFF, and PSNRHMAm, i.e. SSIM and SNR based metrics as well as SFF that employs sparse features.

It is possible to compare the obtained results to other approaches of combined metric design. BMMF [9] provides SROCC equal to 0.834 for TID2013. The best combination of the product of three metrics with the optimized powers gives SROCC=0.8749 [7]. The combined metric that employs trained neural network (NN) [8] provides SROCC about 0.89 but this metric is more complex than the proposed ones. Finally, the most advanced NN-based metrics are characterized by SROCC larger than 0.95 but they are even more complex.

Being optimized or trained for a certain database (TID2013 in the considered case), any metric can perform not well enough for other databases [15-20]. Thus, it is worth verifying the proposed metric performance for other databases.

The most known among them are LIVE Multi-Distortional [15], LIVE [16], and CSIQ [17]. They differ from TID2013 by methodology of choosing test images and experiments performed with volunteers. They are described below:

1) LIVE Multi-Distortional Database consists of 450 test images distorted by combinations of Noise with Blur and JPEG compression with Blur. Each type of distortions has four levels of intensity, where 0 corresponds to no distortion and 1-3 relate to

different levels of intensity according to the predetermined parameters described in [16]. For both pairs of distortions using 15 reference images, 90 images with a single distortion and 135 multiple ones were generated. Image subjective quality assessment was performed using a single stimulus (SS) approach, when a single image (either distorted or reference in random order) was shown to an experiment participant and estimates in the range 1..100 with the semantic labels 'Bad', 'Poor', 'Fair', 'Good' and 'Excellent' were obtained. Difference mean opinion scores (DMOS) for images have been evaluated as the mean of all difference values between reference and distorted images obtained during all experiments. Lower a DMOS, better quality is observed.

2) LIVE is the earlier released database that contains 779 distorted images with single distortions (Gaussian blur, white noise, JPEG compression, JPEG2000 compression and fast fading Rayleigh) that varied in the wide range and were not grouped by predetermined levels. As in the previous database, for subjective experiments a single-stimulus methodology was used "because the number of images to be evaluated was prohibitively large for a double-stimulus study". Experiments were divided to sessions by distortion types. For aggregating results from sessions into one dataset, scale realignment has been made. At the final step, DMOS values with the correction according to realignment were obtained.

3) CSIQ image database contains 30 reference images and 866 distorted images with the most common types of distortions: JPEG compression, JPEG2000 compression, additive white Gaussian noise, additive Gaussian pink noise, Gaussian blur and global contrast decrements. These types of distortion were generated with 4-5 different levels. As authors describe "CSIQ images are subjectively rated based on a linear displacement of the images across four monitors placed side by side on equal viewing distance to the observer. All of the distorted versions of an original image were viewed simultaneously on the monitor array and placed in relation to each other according to an overall quality. Across image ratings are realigned according to a separate, but identical, experiment in which observers place subsets of all the images linearly in space". Ratings are reported in the form of DMOS. For objective comparisons, we further employ fitting functions presented in Table 2 for the considered metrics and use them for the studied databases. The obtained values of SROCC and PC are presented in Table 6.

Comparing the results for elementary metrics for different databases, it is possible to note that a metric performance sufficiently depends on which database it was originally developed and tested. For example, the metrics VIF, NQM, IFC and other metrics developed by A.C. Bovik and his co-authors perform well for databases LIVE and LIVE MD (SROCC values about 0.88 and higher are typical). Meanwhile, for TID2013 their performance is considerably worse. An opposite situation takes place for the metric PSNRHA and its modification PSNRHMA optimized for TID2013 and applied to LIVE MD. The oldest among the considered database LIVE is not "complex" for the best of the modern metrics. Almost all of them have SROCC and PC over 0.9.

Concerning the leaders for CSIQ, they are elementary metrics that have quite low SROCC and PC values for TID2013 whilst they perform rather well for LIVE MD. These results evidence in favor of "overfitting" of metrics that have been trained or optimized for the particular databases. The metrics FSIMc, SFF, SRSIM and MAD_INDEX perform rather well for all four considered databases.

SROCC and PC values for the proposed best metrics are presented in Table 7 for all four databases.

Table 6. SROCC and PC for different databases with fitting according to Table 2

Metrics	TID2013		LIVE MD		LIVE		CSIQ	
	SROCC	PC	SROCC	PC	SROCC	PC	SROCC	PC
PSNR	0.687	0.663	0.677	0.740	0.909	0.911	0.806	0.759
MSSIM	0.787	0.834	0.839	0.874	0.947	0.933	0.915	0.897
SSIM	0.637	0.673	0.642	0.728	0.925	0.937	0.853	0.831
VSNR	0.681	0.661	0.777	0.816	0.939	0.907	0.815	0.758
VIF	0.677	0.767	0.882	0.898	0.976	0.964	0.923	0.928
VIFP	0.608	0.712	0.838	0.871	0.963	0.968	0.888	0.905
NQM	0.635	0.604	0.894	0.897	0.937	0.932	0.736	0.722
WSNR	0.579	0.521	0.763	0.821	0.938	0.940	0.774	0.744
IFC	0.540	0.632	0.885	0.903	0.951	0.956	0.767	0.823
IWSSIM	0.777	0.829	0.883	0.908	0.959	0.936	0.923	0.901
CWSSIM	0.562	0.563	0.621	0.677	0.839	0.814	0.582	0.593
DCTUNE	0.620	0.610	0.810	0.849	0.883	0.898	0.673	0.647
MAD_INDEX	0.781	0.819	0.865	0.894	0.957	0.955	0.947	0.950
PSNRHVS	0.653	0.590	0.713	0.796	0.937	0.918	0.830	0.789
PSNRHVSM	0.624	0.590	0.741	0.818	0.944	0.931	0.822	0.783
PSNRHA	0.819	0.833	0.711	0.780	0.938	0.929	0.926	0.917
PSNRHMA	0.813	0.829	0.740	0.802	0.947	0.934	0.912	0.904
FSIM	0.801	0.857	0.862	0.890	0.961	0.935	0.925	0.909
FSIMc	0.851	0.874	0.867	0.896	0.960	0.938	0.931	0.917
SRSIM	0.807	0.865	0.867	0.887	0.959	0.932	0.932	0.911
SFF	0.851	0.868	0.870	0.887	0.969	0.943	0.963	0.961
PSNRHMAm	0.854	0.856	0.719	0.786	0.938	0.926	0.899	0.877

Table 7. SROCC and PC results of the best combinations for different databases

Databases	Metrics	SROCC	PC
3 metrics (median)			
TID2013	**FSIMc, SFF, PSNRHMAm**	**0.8794**	**0.8964**
LIVE MD	FSIMc, SFF, PSNRHMAm	0.7975	0.8371
LIVE	FSIMc, SFF, PSNRHMAm	0.9627	0.8995
CSIQ	FSIMc, SFF, PSNRHMAm	0.9430	0.9288
5 metrics (median)			
TID2013	**VIFP, DCTUNE, FSIMc, SFF, PSNRHMAm**	**0.8847**	**0.9022**
LIVE MD	VIFP DCTUNE FSIMc SFF PSNRHMAm	0.8507	0.8709
LIVE	VIFP DCTUNE FSIMc SFF PSNRHMAm	0.9620	0.9071
CSIQ	VIFP DCTUNE FSIMc SFF PSNRHMAm	0.9455	0.9419
5 metrics (α-trimmed mean)			
TID2013	**IFC, DCTUNE, FSIMc, SFF, PSNRHMAm**	**0.8871**	**0.9053**
LIVE MD	IFC DCTUNE FSIMc SFF PSNRHMAm	0.8654	0.8885
LIVE	IFC DCTUNE FSIMc SFF PSNRHMAm	0.9631	0.9662
CSIQ	IFC DCTUNE FSIMc SFF PSNRHMAm	0.9457	0.9444

The combined metric based on the median of three elementary metrics (case 1 in Table 3) does not perform well enough for the database LIVE MD. This can be due to bad performance of the elementary metric PSNRHMAm for this case. Meanwhile, this metric performs well for the databases LIVE and CSIQ.

The proposed combined metric that uses median of five estimates of MOS (variant 1 in Table 4) performs better for all four databases. A combination based on α-trimmed mean leads to slightly better results. Thus, the use of five elementary metrics for obtaining MOS estimates via linearization and robust processing of these estimates leads to the positive outcomes. α-trimmed mean (case 1 in Table 5) seems to be the best choice of joint processing. In our opinion, good results are obtained mainly due to universality of the metrics SFF and FSIMc that appear to be "universal" enough for different types of distortions.

We have also used four top-best combined metrics from Table 5 (variants 2…5) for the considered databases Multidistortion-LIVE, LIVE and CSIQ. These metrics do not perform better than the case 1.

Conclusions

Robust linearized combined metrics have been designed for three configurations: median of three estimates of MOS resulting from elementary metrics, median of five estimates, alpha-trimmed mean of five estimates. Fitting and linearization aspects have been considered. Optimization that presumes finding the best sets of elementary metrics among the available ones has been carried out for TID2013 database using one of the criteria –SROCC or standard Pearson correlation. The use of median for MOS of three component metrics has led to SROCC and Pearson correlation coefficient about 0.88 for several good combinations. The use of five component metrics allows further increasing of SROCC and Pearson correlation to 0.9. The proposed solutions have been verified for other existing databases. There are problems remaining for the database LIVE MD although for LIVE and CSIQ databases the proposed metrics perform well.

REFERENCES

[1] W. Lin and C. C. Jay Kuo, "Perceptual visual quality metrics: A survey," Journal of Visual Communication and Image Representation, vol. 22, no. 4, pp. 297-312, 2011.

[2] D. M. Chandler, "Seven Challenges in Image Quality Assessment: Past, Present, and Future Research," ISRN Signal Processing, vol. 2013, pp. 1-53, 2013.

[3] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battistid, and C.-C. Jay Kuo, "Image database TID2013: Peculiarities, results and perspectives," Journal of Signal Processing: Image Communication, vol. 30, pp. 57–77, 2015.

[4] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment, " IEEE Trans. Image Process., vol. 20, no. 5, pp. 2378-2386, 2011.

[5] Okarma, "Colour Image Quality Assessment Using the Combined Full-reference Metric," Computer recognition Systems 4, Advances in Intelligent and Soft Computing, vol. 95, pp. 287-296, 2011.

[6] N. Ponomarenko, O. Ieremeiev, V. Lukin, "Modified Image Visual Quality Metrics for Contrast Change and Mean Shift Accounting," in Proceedings of CADSM, Lviv, Ukraine, 2011.

[7] O. Ieremeiev, V. Lukin, N. Ponomarenko, K. Egiazarian, J. Astola, "Combined full-reference image visual quality metrics," in Proceedings of Image Processing: Algorithms and Systems XIV, San Francisco, 2016.

[8] V. V. Lukin, N. N. Ponomarenko, O. I. Ieremeiev, K. O. Egiazarian and J. Astola, "Combining of full-reference image visual quality metrics by neural network," in Proceedings of SPIE 9394 Human Vision and Electronic Imaging XX, San Francisco, 2015.

[9] L. Jin, K. Egiazarian and C. C. Jay Kuo, "Perceptual image quality assessment using block-based multi-metric fusion BMMF," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012.

[10] T. Liu, K. Liu, J. Lin, W. Lin, and C.-C. J. Kuo, "A paraboost method to image quality assessment," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 1, pp. 107–121, Jan. 2017.

[11] V. Lukin, N. Ponomarenko, K. Egiazarian, J. Astola, Analysis of HVS-Metrics' Properties Using Color Image Database TID2013, Proceedings of ACIVS, October 2015, Italy, pp. 613-624.

[12] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, issue 4, pp. 600-612, 2004.

[13] O. Ieremeiev, V. Lukin, N. Ponomarenko, K. Egiazarian, "Full-reference metrics multidistortional analysis," in Electronic Imaging, Image Processing: Algorithms and Systems XV, pp. 27-35(9), 2017.

[14] H.-W. Chang, H. Yang, Y. Gan, and M.-H. Wang, "Sparse Feature Fidelity for Perceptual Image Quality Assessment," IEEE Trans. Image Process., vol. 22, no. 10, pp. 4007-4018, Oct. 2013. K.

[15] D. Jayaraman, A. Mittal, A. K. Moorthy and A. C. Bovik, "Objective Quality Assessment of Multiply Distorted Images," in Proceedings of Asilomar Conference on Signals, Systems and Computers, Austin, 2012. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_multidistortedimage.html, 5 Nov. 2017.

[16] H.R. Sheikh, Z. Wang, L. Cormack, A.C. Bovik, LIVE Image Quality Assessment Database Release 2, in http://live.ece.utexas.edu/research/quality/subjective.htm, 27 Nov. 2017.

[17] E. C. Larson, D. M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, Journal of Electronic Imaging, 19(2010) 1-21, CSIQ page: http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=23, 7 Nov. 2017.

[18] 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, 10 (2009) 30-45, TID2008 page: http://ponomarenko.info/tid2008.htm, 7 Nov. 2017.

[19] P. Le Callet and F. Autrusseau, "Subjective quality assessment IRCCyN/IVC database," 2005. 2012. [Online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/, 5 Nov. 2017.

[20] Y. Horita, K. Shibata, Z.M. Parvez Saddad, Subjective quality assessment toyama database, in http://mict.eng.u-toyama.ac.jp/mict/, 7 Nov. 2017.