DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

Combined full-reference image visual quality metrics

Oleg I. Ieremeiev[a], Vladimir V. Lukin[a], Nikolay N. Ponomarenko[a], Karen O. Egiazarian[b], Jaakko Astola[b]
[a] National Aerospace University, 61070, Kharkov, Ukraine;
[b] Tampere University of Technology, FIN 33101, Tampere, Finland

Abstract

This paper addresses the problem of assessing full-reference visual quality of images. A correlation between the obtained array of mean opinion scores (MOS) and the corresponding array of given metric values allows characterizing a correspondence of the considered metric to HVS. For the database TID2013 intended for a metric verification, a Spearman correlation is about 0.85 for the best existing HVS-metrics. A simple way to improve an efficiency of assessing visual quality of images is to combine several metrics: as a product of two existing metrics in certain powers that can be optimized or applying more complex structures to unify more than two visual quality metrics. We show that clustering methods can be efficiently used for this purpose. This method provides essentially larger improvement of a combined metric performance compared to the method based on their multiplication. Besides, our work specially addresses assessing images with multiple distortions. There are two such types in the modified LIVE database and two others in TID2013. Spearman rank order correlation coefficient (SROCC) between a combined metric and mean opinion score for a considered database serves as a criterion for the metric optimization. As the result of our design, the SROCC reaches 0.95 for the verification set of the database TID2013. This is considerably better than for any particular metric employed as an input where FSIMc is the best among them.

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

Introduction

Many applications of digital image processing require good full-reference visual quality metrics [1, 2]. Among such applications, it is worth mentioning lossy image and video compression, watermarking, image denoising, etc. Although many (more than 100) full-reference visual quality metrics (indices) have been designed recently and most of them incorporate some heuristics on human visual system (HVS), there is still a need in more adequate universal visual quality metrics as well as in more efficient image quality assessment (IQA) for particular applications.

Design of new HVS-metrics and modification of existing ones deal with several possible approaches. First, some new peculiarities of HVS are taken into account to introduce quite simple modifications into the existing metrics. The examples are the groups PSNR→PSNR-HVS→PSNR-HVS-M→PSNR-HMA [3] and SSIM→MSSIM→FSIM [4]. Second, more complicated metrics are designed with an attempt to use positive features of existing metrics and to avoid their drawbacks. Such metrics can be treated as combined and their examples are the metrics proposed by K. Okarma [5, 6], the metric BMMF [7], and some others [8, 9]. Within these approaches, one important aspect is how complex is a newly designed metric? Another aspect is what is a performance improvement with respect to (elementary, standard) metrics?

To answer the last question, metric verification is usually carried out for one or several existing databases. In particular, these can be LIVE, Toyama, TID2008, TID2013, and others [10-15]. For these databases, the observers (volunteers) have already obtained estimates of image visual quality that have been processed to provide mean opinion score (MOS) to database users. Keeping in mind that HVS-metrics should be in a good correspondence with this MOS, correlation coefficients between the obtained array of MOS and the corresponding array of given metric values are employed to characterize a metric performance (adequateness). Rank correlation coefficients (more often Spearman and less frequently Kendall rank order correlations) have gained popularity for this task where their values approaching to unity correspond to a closer similarity between a metric and human perception of image quality. Note that the aforementioned correlation coefficients are usually determined for images with all types of distortions present in the database to characterize a universality of the metric as well as for a subset or several subsets of distortions to describe an applicability of a given HVS-metric for a particular application where these distortions take place.

It is worth stressing that in practice images are corrupted by multiple distortions. To take this into account, LIVE Multiply Distorted Image Quality Database has been created [16, 17]. Besides, images with two types of multiple distortions are present in TID2013 (such as denoised images and noisy images compressed in a lossy manner).

The main contribution of this paper is the following. First, simple combined metrics which, similarly to those in [5,6], are the products of two standard metrics in a certain (optimized) power, will be considered. Second, the combined metrics that jointly use several standard metrics and employ data clustering principle, will be designed. Third, a special attention will be paid to the cases of multiple distortions present in aforementioned databases. The designed metrics will be compared to the best existing counterparts.

Combining two and three metrics

The idea of joint use of two or more elementary (partial) quality metrics is not new. The goal is to exploit advantages of the used metrics and diminish the influence of their drawbacks. There are several ways to jointly use elementary metrics. First, they can be used as arguments of some function [5, 6]. Second, they can be used as inputs of some 'approximator', e.g., a trained neural network [8]. As it will be shown later, other ways are also possible.

Let us consider the first way of metrics combination. Similarly to the approach of Okarma in [6], let us analyze the combined metrics that can be presented as

$$M_{comb1} = M_1^a \times M_2^b \tag{1}$$

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

where M_1, M_2 are some already known HVS-metrics, a and b are parameters to be optimized.

As metrics used in M_{comb1}, we have chosen those that are among the best for TID2013 [13, 14]. These metrics are the following: FSIM and its color version FSIMc [4], MSSIM [18], SSIM [19], VSNR [20], VIF and VIFP [21], NQM [22], WSNR [23], UQI [24], PSNR-HVS [25] and PSNR-HVS-M [26], PSNR-HMA and PSNR-HA [3], SFF [27], SRSIM [28], IWSSIM [29], IWPSNR [29], and MAD index [30].

As an optimization criterion, Spearman rank order correlation coefficient (SROCC) determined for all 24 types of distortions present in TID2013 between a considered combined metric and mean opinion score (MOS) has been used. The parameters a and b varied in the limits from -2 to +5 with the step 0.05. This allows taking into account such properties of the considered elementary visual quality metrics as non-linearity and different range of possible variation. Note that for all aforementioned metrics larger values correspond to better visual quality.

The optimization results are presented in Table 1. We show only three best pairs for a given metric. For example, for the standard PSNR, the best results were obtained for the metrics SFF, FSIMc, and PSNR-HA. Recall that the best result for one elementary metric is provided by SFF and SROCC for it is equal to 0.8513.

Table 1. Results of optimization for the metric Mcomb1 using all distortions in TID2013

Metric	Combination	SROCC	Combination	SROCC	Combination	SROCC
PSNR	$PSNR^{0.00} \times SFF^{0.05}$	0.8513	$PSNR^{0.00} \times FSIMc^{0.05}$	0.8510	$PSNR^{-0.05} \times PSNRHA^{1.50}$	0.8484
MSSIM	$MSSIM^{1.05} \times SFF^{4.10}$	0.8613	$MSSIM^{-0.20} \times FSIMc^{1.15}$	0.8529	$MSSIM^{3.70} \times PSNRHMA^{0.95}$	0.8335
SSIM	$SSIM^{0.05} \times SFF^{2.55}$	0.8569	$SSIM^{-0.05} \times FSIMc^{4.10}$	0.8517	$SSIM^{0.10} \times PSNRHMA^{0.25}$	0.8298
VSNR	$VSNR^{0.00} \times SFF^{0.05}$	0.8513	$VSNR^{0.00} \times FSIMc^{0.05}$	0.8510	$VSNR^{-0.35} \times PSNRHA^{3.30}$	0.8335
VIF	$VIF^{-0.05} \times FSIMc^{2.55}$	0.8604	$VIF^{0.00} \times SFF^{0.05}$	0.8513	$VIF^{0.15} \times PSNRHMA^{0.85}$	0.8325
VIFP	$VIFP^{0.00} \times SFF^{0.05}$	0.8513	$VIFP^{0.00} \times FSIMc^{0.05}$	0.8510	$VIFP^{0.05} \times PSNRHMA^{0.30}$	0.8356
NQM	$NQM^{0.00} \times SFF^{0.05}$	0.8513	$NQM^{0.00} \times FSIMc^{0.05}$	0.8510	$NQM^{-0.05} \times PSNRHA^{1.60}$	0.8468
WSNR	$WSNR^{0.00} \times SFF^{0.05}$	0.8513	$WSNR^{0.00} \times FSIMc^{0.05}$	0.8510	$WSNR^{-0.05} \times PSNRHA^{1.50}$	0.8478
UQI	$UQI^{0.00} \times SFF^{0.05}$	0.8513	$UQI^{0.00} \times FSIMc^{0.05}$	0.8510	$UQI^{0.00} \times PSNRHA^{0.05}$	0.8187
PSNRHVSM	$PSNRHVSM^{0.00} \times SFF^{0.05}$	0.8513	$PSNRHVSM^{0.00} \times FSIMc^{0.05}$	0.8510	$PSNRHVSM^{-0.15} \times PSNRHA^{4.45}$	0.8488
PSNRHVS	$PSNRHVS^{0.00} \times SFF^{0.05}$	0.8513	$PSNRHVS^{0.00} \times FSIMc^{0.05}$	0.8510	$PSNRHVS^{-0.15} \times PSNRHA^{4.50}$	0.8489
PSNRHMA	$PSNRHMA^{0.05} \times SFF^{1.10}$	0.8601	$PSNRHMA^{0.05} \times FSIMc^{0.90}$	0.8583	$PSNRHMA^{0.10} \times SR\_SIM^{1.35}$	0.8380
PSNRHA	$PSNRHA^{0.05} \times SFF^{0.80}$	0.8630	$PSNRHA^{0.10} \times FSIMc^{1.65}$	0.8569	$PSNRHA^{1.25} \times IWPSNR^{-0.10}$	0.8502
FSIM	$FSIM^{0.75} \times SFF^{1.85}$	0.8633	$FSIM^{-1.30} \times FSIMc^{1.90}$	0.8629	$PSNRHA^{0.25} \times FSIM^{0.95}$	0.8381
FSIMc	**$FSIMc^{0.15} \times SFF^{0.25}$**	**0.8676**	$VIF^{-0.05} \times FSIMc^{2.55}$	0.8604	$PSNRHMA^{0.05} \times FSIMc^{0.90}$	0.8583
SFF	**$FSIMc^{0.15} \times SFF^{0.25}$**	**0.8676**	$SFF^{1.35} \times SRSIM^{1.10}$	0.8661	$FSIM^{0.75} \times SFF^{1.85}$	0.8633
SRSIM	*$SFF^{1.35} \times SRSIM^{1.10}$*	*0.8661*	$FSIMc^{0.35} \times SRSIM^{0.05}$	0.8513	$PSNRHMA^{0.10} \times SRSIM^{1.35}$	0.8380
IWSSIM	$SFF^{0.85} \times IWSSIM^{0.15}$	0.8576	$FSIMc^{1.85} \times IWSSIM^{-0.15}$	0.8521	$PSNRHA^{1.05} \times IWSSIM^{1.70}$	0.8288
IWPSNR	$SFF^{0.05} \times IWPSNR^{0.00}$	0.8513	$FSIMc^{0.05} \times IWPSNR^{0.00}$	0.8510	$PSNRHA^{1.25} \times IWPSNR^{-0.10}$	0.8502
MAD index	$SFF^{0.05} \times MAD\_index^{0.00}$	0.8513	$FSIMc^{0.05} \times MAD\_index^{0.00}$	0.8510	$PSNRHA^{0.05} \times MAD\_index^{0.00}$	0.8187

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

An interesting result has been found for the first combination where PSNR was used as M_1. In the combinations with SFF and FSIMc, the optimal parameter a occurred to be equal to 0 and SROCC is equal to 0.8513. This means that, in fact, PSNR is ignored in the combined metric. Similar results are observed for some other pairs. Besides, an optimal b often occurs to be equal to 0.05. This usually means that the results for any b are the same (e.g., SROCC is the same in the combination $PSNR^{0.00}$ x SFF^{b} for any positive b).

The results which are essentially better than SROCC for the metric SFF are marked by bold and underlined. As it is seen, SROCC for the best combination ($FSIMc^{0.15}$ x $SFF^{0.25}$) reaches 0.8676. This is by 0.016 larger than for SFF used alone. The result is not surprising since both combined metrics possess very good individual performance and jointly contribute to improvement. Quite close result is provided by the metric which is the product $SFF^{1.35}$ x $SRSIM^{1.10}$. This shows that the metrics used as elementary ones in the combined metric should be both good. Desirably, they should belong to different groups to incorporate different features of HVS.

Consider now a more sophisticated combined metric (the second type of combined metrics) constructed as

$$M_{comb2} = M_1^a \times M_2^b \times M_3^c \quad (2)$$

where c is the real-valued parameter. The optimization criterion and the methodology was the same. Since now we have more varied parameters, the step of the parameter variation is larger. Only the best two combinations are shown in each line of Table 2.

The maximal obtained SROCC is equal to 0.8744 which is slightly larger than that for the best M_{comb1}. This SROCC is observed for the combined metric $FSIM^{-1.40}$ x $FSIMc^{1.30}$ x $SR\_SIM^{0.80}$. This result is quite interesting since both variants of the metric FSIM (grayscale FSIM and the color version FSIMc) participate in it and they are in negative (-1.4) and positive (1.3) powers, respectively.

Quite good combinations are also $FSIM^{-1.10}$ x $FSIMc^{1.40}$ x $SFF^{0.60}$ and $PSNRHA^{0.20}$ x $FSIMc^{2.10}$ x $SFF^{4.50}$ both providing SROCC of the combined metric with MOS over 0.87. Meanwhile, optimization often produced results as, e.g., $FSIMc^{1.50}$ x $SFF^{2.50}$ x $IWPSNR^{0.00}$ which is, in fact, equivalent to the combined metric of the first type $FSIMc^{1.50}$ x $SFF^{2.50}$.

Therefore, a preliminary conclusion is the following: even simple combinations of elementary metrics can produce performance improvement, but a larger number of used elementary metrics provide better results.

Table 2. Results of optimization for the metric using all distortions in TID2013

Metric	Combination	SROCC	Combination	SROCC
PSNR	$PSNR^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$PSNR^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
MSSIM	$MSSIM^{-0.30}$ x $FSIMc^{3.10}$ x $SFF^{4.50}$	0.8678	$MSSIM^{0.20}$ x $SFF^{3.10}$ x $SRSIM^{2.30}$	0.8662
SSIM	$SSIM^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$SSIM^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
VSNR	$VSNR^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$VSNR^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
VIF	$VIF^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$VIF^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
VIFP	$VIFP^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$VIFP^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
NQM	$NQM^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$NQM^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
WSNR	$WSNR^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$WSNR^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
UQI	$UQI^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$UQI^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
PSNRHVSM	$PSNRHVSM^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$PSNRHVSM^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
PSNRHVS	$PSNRHVS^{0.00}$ x $FSIMc^{1.50}$ x $SFF^{2.50}$	0.8676	$PSNRHVS^{0.00}$ x $SFF^{3.00}$ x $SRSIM^{2.50}$	0.8661
PSNRHMA	$PSNRHMA^{0.10}$ x $FSIMc^{1.90}$ x $SFF^{3.70}$	0.8695	$PSNRHMA^{0.10}$ x $SFF^{4.90}$ x $SR\_SIM^{3.90}$	0.8678
PSNRHA	$PSNRHA^{0.20}$ x $FSIMc^{2.10}$ x $SFF^{4.50}$	0.8701	$PSNRHA^{0.10}$ x $SFF^{3.10}$ x $SRSIM^{2.00}$	0.8681
FSIM	**$FSIM^{-1.40}$ x $FSIMc^{1.30}$ x $SR\_SIM^{0.80}$**	**0.8749**	*$FSIM^{1.10}$ x $FSIMc^{1.40}$ x $SFF^{0.60}$*	*0.8717*
FSIMc	**$FSIM^{-1.40}$ x $FSIMc^{1.30}$ x $SR\_SIM^{0.80}$**	**0.8749**	*$FSIM^{1.10}$ x $FSIMc^{1.40}$ x $SFF^{0.60}$*	*0.8717*
SFF	*$FSIM^{1.10}$ x $FSIMc^{1.40}$ x $SFF^{0.60}$*	*0.8717*	$PSNRHA^{0.20}$ x $FSIMc^{2.10}$ x $SFF^{4.50}$	0.8701
SRSIM	**$FSIM^{-1.40}$ x $FSIMc^{1.30}$ x $SR\_SIM^{0.80}$**	**0.8749**	$FSIMc^{2.20}$ x $SFF^{4.70}$ x $SRSIM^{1.60}$	0.8683
IWSSIM	$FSIMc^{3.10}$ x $SFF^{4.20}$ x $IWSSIM^{-0.50}$	0.8684	$SFF^{4.10}$ x $SRSIM^{4.90}$ x $IWSSIM^{-0.60}$	0.8674
IWPSNR	$FSIMc^{1.50}$ x $SFF^{2.50}$ x $IWPSNR^{0.00}$	0.8676	$SFF^{3.00}$ x $SRSIM^{2.50}$ x $IWPSNR^{0.00}$	0.8661
MAD index	$FSIMc^{1.50}$ x $SFF^{2.50}$ x MAD index $^{0.00}$	0.8676	$SFF^{3.00}$ x $SRSIM^{2.50}$ x MAD index $^{0.00}$	0.8661

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

Combining several metrics using clustering

Alongside with the combined metrics M_{comb1} and M_{comb2}, we have proposed to design the combined metrics based on the clustering. It allows using more than two or three particular metrics. We have analyzed the cases from 2 to 8 metrics. Let us keep in mind that more metrics means more possible combinations at the design stage and more calculations at the stage when combined metrics are used.For clustering based combined metrics, the leaning stage was needed similarly to NN-based metrics [8]. Both learning and verification have been carried out for the database TID2013 divided into equal parts (1500 images used for training and 1500 for verification). Numerous combination sets of the particular metrics chosen from the best 20 metrics randomly have been exploited. A combined metric for a given image is obtained as the result of corresponding particular metric comparison with its threshold within the clustering tree till falling a final cluster and assigning the combined metric value to it. This value is obtained as an average MOS for a given cluster. Factors that influence clustering efficiency are considered below.

There are numerous clustering techniques. Here we have employed a popular method of k-means [31, 32]. For a good operation of this method (correct calculation of cluster centers), it is desirable to provide that set element coordinates have equal scale and have linear variation scale. To ensure this, for all metrics used in clustering, robust fitting of these metrics to MOS for the database TID2013 has been done using the power functions as:

$$y = d * x^e + f, \quad (3)$$

where x is a metric, y denotes a fitted metric value in MOS range, d, e, and f denote the coefficients obtained as a result of fitting using standard Matlab tools (bisquare).A list of elementary metrics used in the cluster-based approach to combined metric design and the obtained values of parameters d, e, and f are presented in Table 3.

Then, we have carried out pair-wise clustering of these metrics (SFF / PSNRHMA, FSIMc / PSNRHVSM, SRIM / VIF, VSNR / PSNRHA). For training we have used a half (1500) of MOS of the database TID2013 chosen from entire array of MOS values randomly ("learning or training) set"). The remained 1500 MOS values have been used for the verification of clustering quality, i.e., as a "verification set").

Table 3. Metrics and results of fitting

N	Metric	D	e	f
1	SFF	-3.981	18.23	1.82
2	PSNRHMA	-265.2	-1.194	8.643
3	FSIMc	3.751	10.63	2.06
4	PSNRHVSM	-1129	-1.916	6.128
5	SRIM	3.806	19.8	1.92
6	VIF	5.765	0.3881	-0.07862
7	VSNR	-46.18	-0.9676	6.608
8	PSNRHA	-526.4	-1.444	8.326

Entire set of elements has been distributed between 25 clusters for each metric after fitting. If some cluster had less than 10 elements (the influence of cluster number and minimal size will be discussed below), then elements of this cluster are spread between larger neighbor clusters. Fig. 1 presents clustering results for the elementary metrics SFF and PSNRHMA.

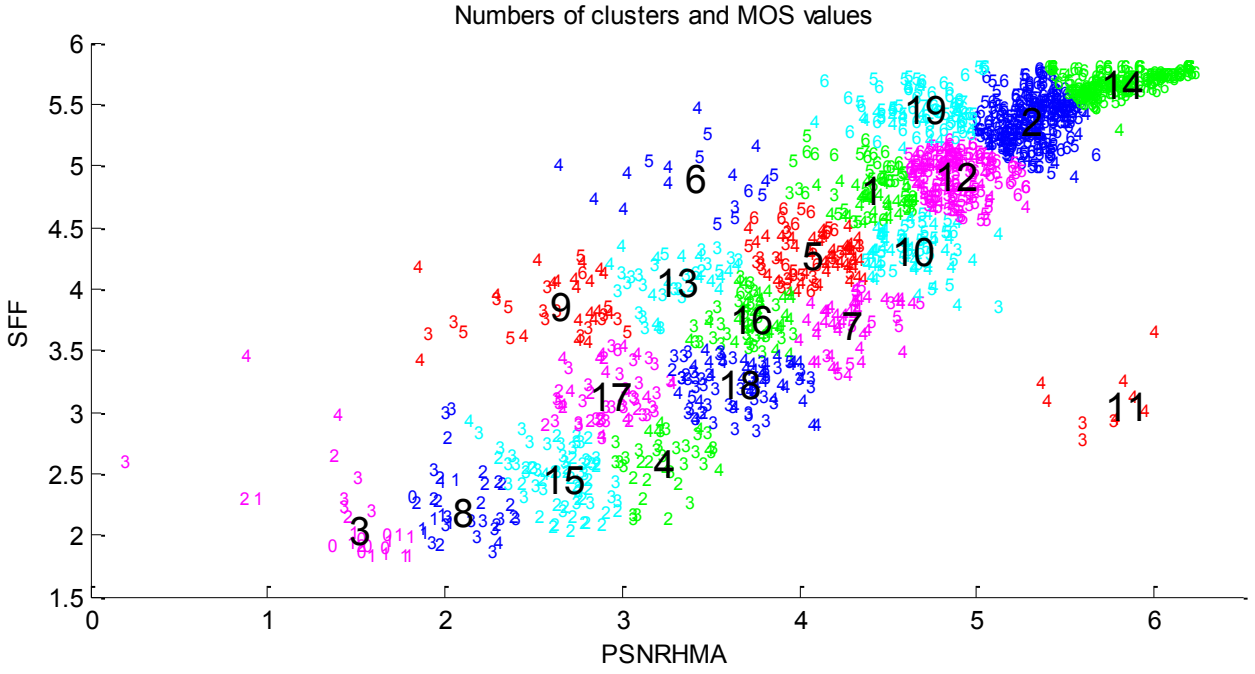

Figure 1. Result of clustering the metrics PSNRHMA and SFF

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

Large black color digits show indices (numbers) of clusters where we have got 19 clusters after reforming small size clusters. Small digits show MOS for each element (particular image in the database TID2013). MOS of images that fell into the same cluster is shown by the same color to differ elements of neighbor clusters. Note that such representation is useful in analysis of metric drawbacks. For example, for clustering in Fig. 1, there are such clusters found for which the results for one metric are obviously incorrect. These are the results for the metric PSNRHMA for cluster #11, the results for the metric SFF for cluster #7. There are also clusters where there is a weak correspondence between MOS and both metrics (e.g., cluster #3) and clusters where there is a strong correspondence for both metrics and they can support each other (cluster #2).

Then, for elements of each cluster and both metrics, their fitting to MOS is accomplished using a first order polynomial:

$$y = g * x + h, \quad (4)$$

where x is a metric value for the cluster elements, y denotes a forecasted MOS value, g and h are coefficients obtained by the standard least mean square error (LMSE) fitting. For each cluster and both metrics, the parameters g and h are saved (denote them as g_1, h_1, g_2, h_2) for the considered two metrics. The fitting errors characterized by fitting RMSE (denote them as σ_1 and σ_2) are saved as well.Then, for the combined metric obtained by clustering its value is determined according to the following algorithm:

1. Calculate the values of the first and second used metrics m_1 and m_2 for a given image.
2. Using expressions (3) and parameters given in Table 3, calculate the values of these metrics transformed by fitting into MOS scale. Denote these new values as m_{m1} and m_{m2}.
3. Find the cluster which is the closest to m_{m1} and m_{m2}.
4. According to expression (4) and the saved (for the determined cluster) values g_1, h_1, g_2, h_2, σ_1 and σ_2, calculate the resulting weighted value of the combined (integral) metric m_i as

$$m_i = ((m_{m1}g_1 + h_1)/\sigma_1 + (m_{m2}g_2 + h_2)/\sigma_2) / (\sigma_1 + \sigma_2). \quad (5)$$

Note that instead of weighted averaging (5) it is also possible to set m_i equal to m_{m1}, if $\sigma_1 < \sigma_2$ and m_{m2} otherwise.

Note that, on one hand, a larger number of clusters and a smaller limit on the minimal acceptable cluster size leads to more accurate estimation of MOS according to (5) for the training set. On the other hand, it is then difficult to accurately estimate parameters used in expressions (4) and (5) which results in a less accurate MOS estimation for the verification set and for other practical images (which are not in TID2013). To have some trade-off, one can use as initial number of clusters approximately 25 and a minimal number of elements in small clusters as 0.5%...1% of the total number of elements in the training set.

After pairwise clustering of elementary metrics, we have carried out pairwise clustering of the obtained integral metrics. As a result, two sets of clustering results have been obtained where the first set takes into account the following metrics: SFF, PSNRHMA, FSIMc, and PSNRHVSM, and the second set accounts for the metrics SRIM, VIF, VSNR, and PSNRHA. Finally, at the final stage of clustering, all eight metrics have been combined in the final integral metric.

Table 4 presents SROCC values for all stages of clustering. Besides, for comparison, we give SROCC values for some elementary and combined metrics, presented earlier in Tables 1 и 2, for both sets. This allows us to analyze a variability of SROCC depending on the used subset. Note that "training" and "verification" have no meaning for the combined metrics.

Table 4. SROCC values for training and verifications sets

Metric	Training subset	Verification subset
Separate metrics		
FSIMc	0.845	0.857
SFF	0.850	0.853
PSNRHMA	0.809	0.817
PSNRHVSM	0.604	0.646
Metrics combined according to (1) and (2)		
$PSNRHMA^{0.05}SFF^{1.10}$	0.857	0.864
$PSNRHMA^{0.10}\,SFF^{3.70}\,FSIMc^{1.90}$	0.866	0.874
$FSIM^{-1.40}\,FSIMc^{1.30}\,SRSIM^{0.80}$	0.871	0.879
Metrics combined by clustering of 2 metrics		
PSNRHMA / SFF	0.867	0.864
PSNRHVSM / FSIMc	0.873	0.871
SRSIM / VIF	0.862	0.874
VSNR / PSNRHA	0.859	0.859
Metrics combined by clustering of 4 metrics		
PSNRHMA / SFF / PSNR-HVS-M / FSIMc	0.895	0.880
SRSIM / VIF / VSNR / PSNRHA	0.898	0.896
Metric combined by clustering of 8 metrics		
PSNRHMA / SFF / PSNR-HVS-M / FSIMc / SRSIM / VIF / VSNR / PSNRHA	**0.916**	**0.901**

The analysis of data in Table 4 shows the following. First, imperfectness of clustering results in less SROCC for the verification set than that for the training set of all clustering based combined metrics. Second, more elementary metrics are used in clustering combined metric, larger SROCC is provided where the largest value exceeds 0.9 for both training and verifications sets.

Data for some combined metrics obtained according to (1) show that some combinations do not produce benefits compared to the better elementary metric (see the results for the metric $PSNRHVSM^{0.00}$ x $FSIMc^{0.05}$). However, clustering based approach allows reaching SROCC=0.873 for training and 0.871 for verification sets, i.e., better than for any combined metric of the type (1). Recall, that SROCC for FSIMc is equal to 0.857 and SROCC for PSNRHVSM is only 0.646.

Analysis of data in Table 4 shows that the best positive effect due to clustering for two elementary metrics is observed if one metric is suited well for taking into account color distortions while another metric is designed for grayscale images.

Fig. 2 presents the scatter-plot of MOS for the metric FSIMc which has the largest SROCC for elementary metrics considered in this paper. For the convenience of analysis, different levels of distortions in TID2013 are presented in the scatter-plot by different colors (the 1-st (smallest) level by green, the 2-nd level by cyan,

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

the 3-rd level by blue, the 4-th level by red and the 5-th (highest) level by black (note that MOS values in TID2013 are from almost zero to approximately 7.2). The integer numbers from 1 to 24 in the scatter-plot correspond to distortion type indices in TID2013 (totally, 24 types). Small letters of Latin alphabet from *a* to *y* near numbers correspond to indices of reference images in the database from 1 to 25. Then, each point of the scatterplot can be easily connected with the corresponding distorted image in TID2013. For example, black mark 13b corresponds to the 5-th level of distortions for the 13-th distortion type for the second reference image (the letter b is at the second position in the alphabet); this corresponds to the image i02_13_5.bmp of the database TID2013.

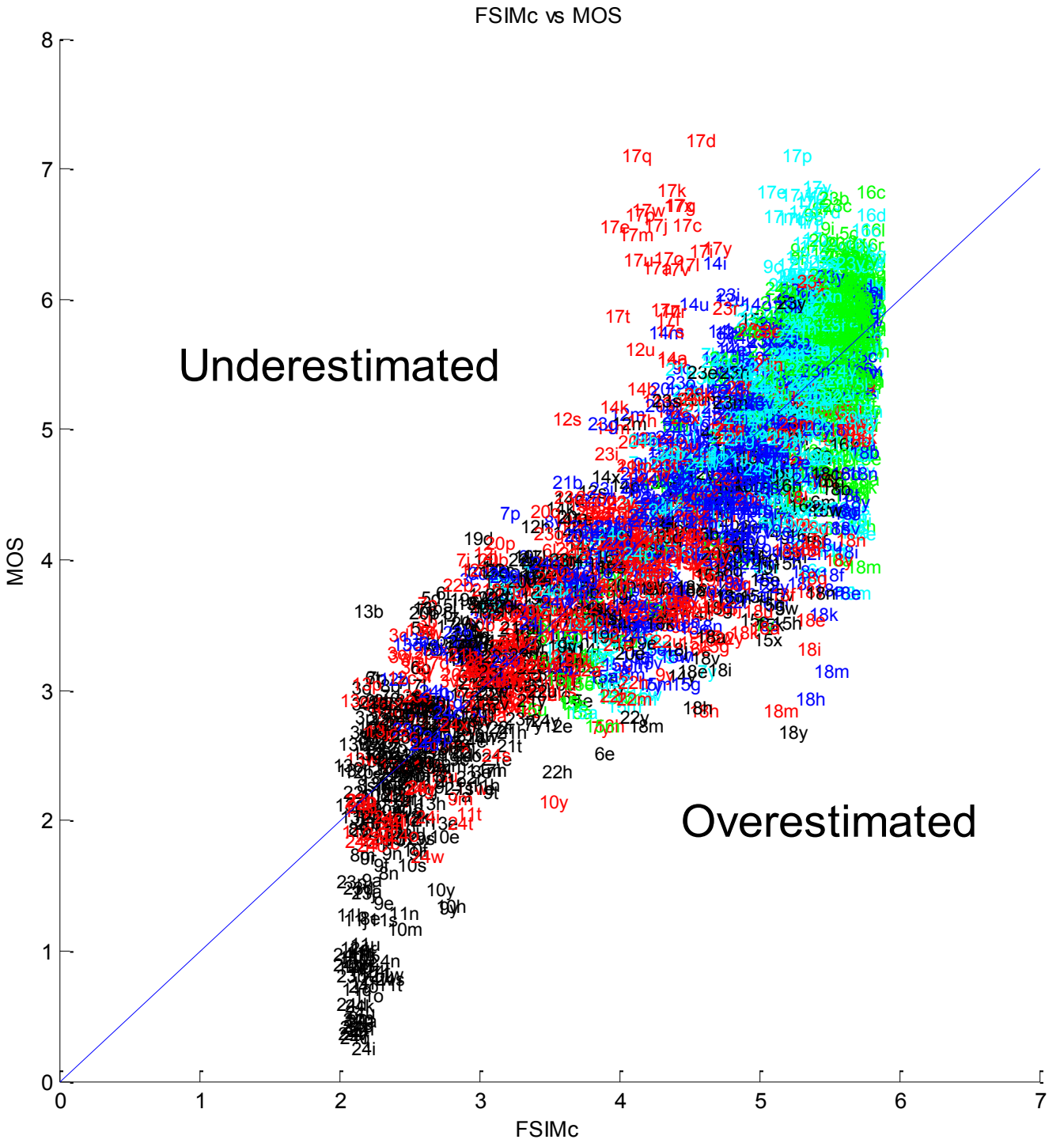

Figure 2. Scatter-plot of MOS values for FSIMc metric

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

Analysis of the scatter-plot in Fig. 2 shows that the metric FSIMc sufficiently underestimates visual quality for the 17-th type of distortions (contrast changes) and considerably overestimates quality for the 18-th type of distortions (change of color saturation) and most images with distortion level #5 (that have low visual quality). Meanwhile, the metric demonstrates good results for most types of distortions for relatively low levels of distortions (from the first to the third level).

The scatter-plot for the combined clustering based metric (8 elementary metrics) is shown in Fig. 3.

It is seen well that a quality of evaluation has radically increased for the 17-th and 18-th types of distortions and for the fifth level of distortions. However, the correspondence between the metric values of MOS is still worth improving.

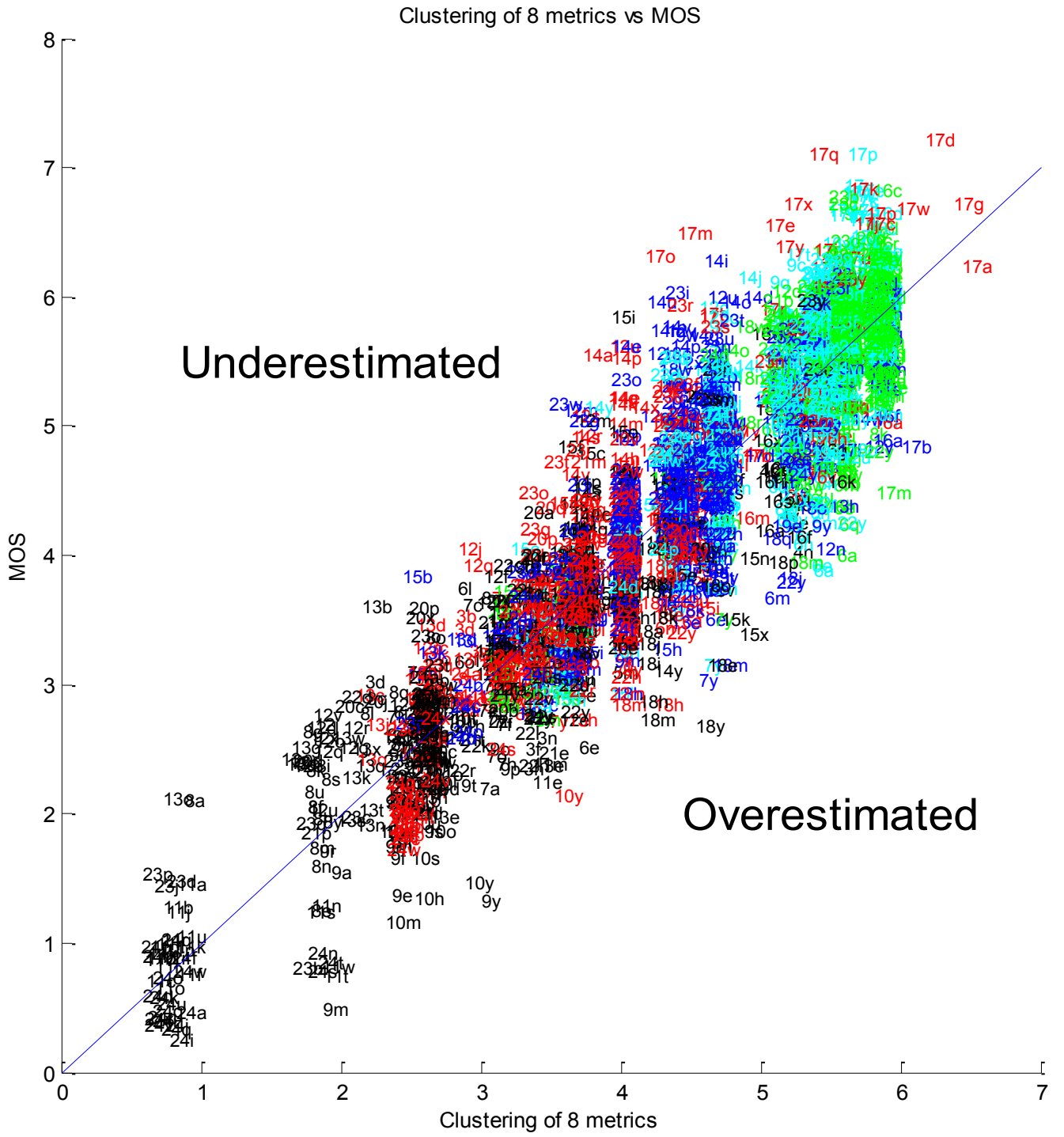

Figure 3. Scatter-plot of MOS vs the designed combined clustering based metric that uses 8 elementary metrics

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

Performance analysis for multiple distortions

The designed combined metrics M_{comb1} and M_{comb2} have been tested for three cases:

- images with multiple distortions (md) in the LIVE Multiply Distorted Image Quality Database (further denoted as MDLIVE;
- all images in MDLIVE;
- multiple distortions (## 9 and 21) in TID2013.

The obtained data are presented in Table 4. If it is written in column 2 "only SFF", it means that for M_{comb1} a=0 in (1). Similarly, if it is written in column 6 "only SFF and FSIMc", it also means that a=0 in (2).

First, we would like to present the results for the best three- and four-component combined metrics proposed in [6] and applied to images in TID2013. For three-component combined metric (based on VSNR, NQM, and IFC), the SROCC for all 24 types of distortions is equal to 0.7063. This is quite low value taking into account that SROCC for elementary metric SFF is 0.8513. SROCC for multiple distortions is 0.9027. This is not bad, however, at the same time, not excellent (see data in Table 4 and analysis below).

Similarly, for four-component combined metric proposed in [6] (based on VSNR, NQM, IFC, and VIF), the SROCC for all 24 types of distortions is equal to 0.7061 (again low) and SROCC for multiple distortions is equal to 0.9153. Thus, the optimization results obtained in [6] for the database MDLIVE do not perform too well for the other databases (in particular, for TID2013).

Data for images with multiple distortions (md) in the LIVE Multiply Distorted Image Quality Database (the first case) allow comparing the results of the combined metric optimization carried out for one database (TID2013) with the results for the other database. The best elementary metric SFF (according to analysis for TID2013) produces SROCC equal only to 0.761, i.e. quite low. The combined metric M_{comb1} based on FSIMc and SFF (see Section 2) produces SROCC equal to 0.7653. The best M_{comb1} occurs to be based on IWSSIM and SFF and SROCC for it is equal to 0.7752. The results for the second type of combined metric (M_{comb2}) are even worse (see data in Table 4). SROCC in the best case is slightly larger than 0.77 and it is not enough. This shows that the optimization results for different databases differ a lot. We have applied the designed combined clustering based metric. The SROCC value for it is 0.74. Thus no benefit is provided.

Table 5. SROCC values for different combined metrics

Metric	Combined with	M_{comb1}			Combined with	M_{comb2}		
		MDLIVE (md)	MDLIVE (Full)	TID2013 (9,21)		MDLIVE (md)	MDLIVE (Full)	TID2013 (9,21)
PSNR	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
MSSIM	SFF	0.7478	0.8658	0.9311	SFF and FSIMc	0.7640	0.8719	0.9413
SSIM	SFF	0.7532	0.8673	0.9261	only SFF and FSIMc	0.7653	0.8725	0.9407
VSNR	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
VIF	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
VIFP	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
NQM	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
WSNR	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
UQI	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
PSNRHVSM	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
PSNRHVS	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
PSNRHMA	SFF	0.7374	0.8575	0.9427	SFF and FSIMc	0.7591	0.8692	0.9447
PSNRHA	SFF	0.7238	0.8506	**0.9493**	SFF and FSIMc	0.7530	0.8657	0.9492
FSIM	SFF	0.7645	0.8713	0.9367	FSIMc and SRSIM	0.7705	0.8766	**0.9564**
FSIMc	SFF	0.7653	0.8725	0.9407	FSIM and SRSIM	0.7705	0.8766	**0.9564**
SFF	FSIMc	0.7653	0.8725	0.9407	FSIM and FSIMc	**0.7708**	**0.8769**	0.9430
SRSIM	SFF	0.7684	0.8727	0.9436	FSIM and FSIMc	0.7705	0.8766	**0.9564**
IWSSIM	SFF	**0.7752**	**0.8780**	0.9295	SFF and FSIMc	0.7572	0.8676	0.9421
IWPSNR	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407
MAD index	only SFF	0.7610	0.8699	0.9257	only SFF and FSIMc	0.7653	0.8725	0.9407

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

Data for the case two (all images in MDLIVE) demonstrate that all presented combined metrics perform for MDLIVE well enough. SROCC for SFF is 0.8699, i.e. larger than for all types of distortions in TID2013. The metric M_{comb1} based on FSIMc and SFF produces SROCC equal to 0.8725. The best M_{comb1} is again based on IWSSIM and SFF and SROCC for it is equal to 0.8780. These are quite high SROCC values. The best results for M_{comb2} are at the same level (see data in Table 5). The largest SROCC is observed for elementary metrics SFF, FSIM and FSIMc (0.8769). Note that M_{comb1} and M_{comb2} have not been re-optimized to MDLIVE. The SROCC value for the best combined clustering based metric is 0.86. This is approximately at the same level as the best elementary metric.

Finally, for the case 3 (multiple distortions in TID2013), all combined metrics perform well providing SROCC over 0.92 for M_{comb1} and over 0.94 for M_{comb2}. For M_{comb1}, the best result is provided by the combined metric on basis of PSNR-HA and SFF (SROCC=0.9493). In turn, for $M_{comb2,}$ the best performance is observed for the metric on basis of elementary metrics SRSIM, FSIM and FSIMc (SROCC=0.9564). The best clustering-based combined metric produces SROCC exceeding 0.95, i.e. at the same level with the best metrics.

CONCLUSIONS

We have considered two ways of combining several full-reference visual quality metrics. The general tendency is that by increasing the number of particular (input) metrics it is possible to provide better performance characterized by SROCC calculated between a combined metric and MOS. Such a combination can be helpful for a metric universality (increasing SROCC for all types of distortions) and for particular applications where multiple distortions can be met. Optimization of a combined metric performed for one database can be not efficient for other databases. Clustering-based combined metric produces better results than simple combinations of elementary metrics.

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, O. Ieremeiev, V. Lukin and K. Egiazarian, "Modified image visual quality metrics for contrast change and mean shift accounting," in 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Polyana-Svalyava, Ukraine, 2011.

[4] L. Zhang, L. Zhang, X. Mou and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," IEEE Transactions on Image Processing, vol. 20, no. 8, pp. 2378-2386, 2011.

[5] K. 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] K. Okarma, "Quality Assessment of Images with Multiple Distortions Using Combined Metrics," *Elektronika ir Elektrotechnika,* vol. 20, no. 6, pp. 128-131, 2014.

[7] 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.

[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] R. Gupta, D. Bansal and C. Singh, "Image Quality Assessment Using Non-Linear MultiMetric Fusion Approach," International Journal of Emerging Technology and Advanced Engineering, vol. 4, no. 7, pp. 822-826, 2014.

[10] H. Sheikh, Z. Wang, L. Cormack and A. Bovik, "LIVE Image Quality Assessment Database Release 2," [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm. [Accessed 25 Nov. 2015].

[11] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli and F. Battisti, "TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics," Advances of Modern Radioelectronics, vol. 10, no. 4, pp. 30-45, 2009.

[12] Y. Horita, K. Shibata and Y. Kawayoke, "Toyama Image Quality Evaluation Database," [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html. [Accessed 25 Nov. 2015].

[13] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian and others, "Color image database TID2013: Peculiarities and preliminary results," in 4th European Workshop on Visual Information Processing (EUVIP), Paris, 2013.

[14] N. Ponomarenko, O. Ieremeiev, V. Lukin, L. Jin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli and F. Battisti, "A New Color Image Database TID2013: Innovations and Results," Advanced Concepts for Intelligent Vision Systems, vol. 8192, pp. 402-413, 2013.

[15] S. Winkler, "Analysis of Public Image and Video Databases for Quality Assessment," IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, pp. 616-625, 2012.

[16] D. Jayaraman, A. Mittal, A. K. Moorthy and A. C. Bovik, "LIVE Multiply Distorted Image Quality Database," 2012. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_multidistortedimage.html. [Accessed 25 Nov. 2015].

[17] 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.

[18] Z. Wang, E. P. Simoncelli, A.C. Bovik, "Multiscale structural similarity for image quality assessment, " in Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 2004.

[19] Z. Wang, A. C. Bovik, H.R. Sheikh, E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity, " IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

[20] D. M. Chandler, S. S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images, " IEEE Transactions on Image Processing, vol. 16, no .9, pp. 2284-2298, 2007.

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-180

[21] H. R. Sheikh, A. C. Bovik, "Image information and visual quality," IEEE Transactions on Image Processing, vol. 15,no. 2, pp. 430-444, 2006.

[22] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, A. C. Bovik, "Image quality assessment based on a degradation model," IEEE Transactions on Image Processing, vol. 9,no. 8, pp. 636-650, 2000.

[23] T. Mitsa, K. L. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Minneapolis, USA, 1993.

[24] Z. Wang, A. C. Bovik, "A universal image quality index," IEEE Signal Processing Letters, vol. 9, no. 3, pp. 81-84, 2002.

[25] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, M. Carli, "New full-reference quality metrics based on HVS," in Second International Workshop on Video Processing and Quality Metrics, Scottsdale, USA, 2006.

[26] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, V. Lukin,"On between-coefficient contrast masking of DCT basis functions," in Third International Workshop on Video Processing and Quality Metrics, Scottsdale, USA, 2007.

[27] H.-W. Chang, H. Yang, Y. Gan and M.-H. Wang, "Sparse Feature Fidelity for Perceptual Image Quality Assessment," IEEE Transactions on Image Processing, vol. 22, no. 10, pp. 4007-4018, 2013.

[28] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in 19th IEEE International Conference on Image Processing (ICIP), Orlando, USA, 2012.

[29] Z. Wang and Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment," IEEE Transactions on Image Processing, vol. 20, no. 5, pp. 1185-1198, 2011.

[30] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," Journal of Electronic Imaging, Special Section on Image Quality, vol. 19, no. 1, pp. 011006-011021, 2010.

[31] K. Jajuga, A. Sokolowski and H. H. Bock, Classification, Clustering, and Data Analysis: Recent Advances and Applications (Studies in Classification, Data Analysis, and Knowledge Organization), Springer, New York, 2002.

[32] R. Xu and D. C. Wunsch, Clustering, Wiley-IEEE Press, New Jersey, 2008.

[33] D. MacKay, Information Theory, Inference and Learning Algorithms, Cambridge Unversity Press, New York, 2003.

Author Biography

Oleg Ieremeiev received his MSc in Telecommunications in 2009 and his diploma of Candidate of Science (comparable to PhD) in Telecommunications in 2015 from the National Aerospace University in Ukraine. Since 2010 he has worked as researcher in Department of Signal Reception, Transmission and Processing of National Aerospace University. His work has focused on image processing and visual quality assessment.