https://doi.org/10.2352/ISSN.2470-1173.2017.13.IPAS-202

Full-reference metrics multidistortional analysis

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

This paper is devoted to analysis and further improvement of full-reference metrics of image visual quality. The effectiveness of a metric is characterized by the rank correlation factors between the obtained array of mean opinion scores (MOS) and the corresponding array of given metric values. This allows to determine the correspondence of a considered metric to a human visual system (HVS). Results obtained on the database TID2013 show that Spearman correlation for the best existing metrics (PSNRHMA, FSIM, SFF, etc.) does not exceed 0.85. In this paper, extended verification tools that allow to detect the shortcomings of the metrics taking into account combined distortions is proposed. An example for further improvement of the PSNRHMA metric is presented.

Keywords: image visual quality assessment, full-reference metrics, metrics verification, multiple distortions, metrics analysis.

Introduction

Fast developments of information technologies leads to a considerable expansion of the areas of applying digital images and image quality assessment (IQA) [1, 2]. Visual quality metrics have become very useful in image processing and analysis including such applications as lossy compression, denoising, watermarking, deblurring, classification, object detection, content based image retrieval, etc. To provide a metric's adequacy, peculiarities of human visual system (HVS) are often taken into account in one or another manner [2, 3].

Although tens of different visual quality metrics have been proposed (see, e.g., [2, 3] and references therein), their performance is still worth improving. For example, a verification of more than twenty modern visual quality metrics for the largest openly available database of distorted test images TID2013 [3] has demonstrated that even the best metrics (FSIMc [4], SFF [5], PSNR-HMA [6]) produce Spearman rank order correlation coefficient (SROCC) with mean opinion score (MOS) of the order 0.85 and less. This shows that the above mentioned metrics are not universal enough (although there are certain types of distortions for which metrics adequacy is high enough). This makes desirable to further improve metrics' performance. Then, the metrics drawbacks have to be detected first and this is one of the goals of this paper.

Note that there are several ways to solve this task. Researchers consider and incorporate more sophisticated models of HVS, design combined metrics [7, 8] and/or employ learning techniques for artificial neural networks etc. [9, 10]. Another approach consists in creating new sets of test images (databases) and obtaining the judgments from observers for distorted images [3, 11-15]. Further improvement of metrics' performance can be prevented by low accuracy of assessments for databases (if amount of experiments was not high enough) and/or if some types of distortions have not been taken into account.

One type of distortions that have attracted attention recently can be treated as combined or multiple ones [16]. Really, almost all databases contain images with particular types of distortions as, e.g., additive white Gaussian noise (AWGN), blur, distortions due to JPEG or JPEG2000 and so on. Meanwhile, real life images are frequently corrupted by multiple (combined) distortions. For instance, digital image acquired in a condition of bad illumination can suffer from blur due to incorrect focusing, noise, and compression artifacts (other types can be present as well). Depending on a type of the camera and conditions of image acquisition, these distortions can appear themselves in different ways. For example, DSLR cameras with good sensors suffer from noise in less degree but distortions due to optics can become prevailing. Lower cost digital cameras mounted in mobile phones, smartphones, notebooks, might produce images with different combinations of aforementioned types of distortions. To provide an efficient and adequate assessment of visual quality for such images, one needs to have HVS-metric(s) that are able to perform well for such multiple distortions.

To partly alleviate this problem, LIVE Multiply Distorted Image Quality Database (LIVE MD) [16, 17] has been created recently that contains images with multiple distortions. The database TID2013 also has a few sets of images with multiple distortions. Analysis of metrics performance for such images is the second task we deal in this paper. Based on the obtained results, we show how a particular metric (PSNR-HMA) can be modified.

Limitations of existing databases

Accuracy of verification of quality metrics sufficiently depends on several factors and primarily on the choice of a database of distorted images. Sometimes, to solve a particular task, it might be enough to have a small number of test images with a few types of distortions. However, to design an accurate and universal visual quality metric, one needs a database that contain images with various distortion types typical for practice where MOS values are derived for sufficiently large number of experiments carried out by observers. These requirements are satisfied by existing databases in larger or less degree (comparison of some databases is given in [3]). However, combined (multiple) distortions are not well represented in the existing databases.

One option to reach our goal is to use the largest available database TID2013. Recall that this database contains 25 reference (distortion-free) color images of equal size where 24 images were obtained (by cropping) from the Kodak database http://r0k.us/graphics/kodak/. The 25-th reference image was artificially created and added to 24 natural scene images with the aim to analyze applicability of metrics to characterize a quality of an artificial image. In fact, TID2013 is a considerable modification of the database TID2008 [14]. TID2013 contains images with 24 types of distortions given in Table 1. TID2008 contained 17 types of distortions and seven new ones with indices 18-24 were added to better represent color distortions (## 18, 22, 23) or new emerging applications for which analysis had not been done yet (## 19-22, 24).

Table 1. List of distortion types in TID2013 and used subsets

#	Type of distortion (four levels for each distortion)	MD	Noise	Actual	Simple	Exotic	Color	Full
1	Additive Gaussian noise	-	+	+	+	-	-	+
2	Additive noise mostly in color components	-	+	-	-	-	+	+
3	Spatially correlated noise	-	+	+	-	-	-	+
4	Masked noise	-	+	+	-	-	-	+
5	High frequency noise	-	+	+	-	-	-	+
6	Impulse noise	-	+	+	-	-	-	+
7	Quantization noise	-	+	-	-	-	+	+
8	Gaussian blur	-	+	+	+	-	-	+
9	Image denoising	+	+	+	-	-	-	+
10	JPEG compression	-	-	+	+	-	+	+
11	JPEG2000 compression	-	-	+	-	-	-	+
12	JPEG transmission errors	-	-	-	-	+	-	+
13	JPEG2000 transmission errors	-	-	-	-	+	-	+
14	Non eccentricity pattern noise	-	-	-	-	+	-	+
15	Local block-wise distortions of different intensity	-	-	-	-	+	-	+
16	Mean shift (intensity shift)	-	-	-	-	+	-	+
17	Contrast change	-	-	-	-	+	-	+
18	Change of color saturation	-	-	-	-	-	+	+
19	Multiplicative Gaussian noise	-	+	+	-	-	-	+
20	Comfort noise	-	-	-	-	+	-	+
21	Lossy compression of noisy images	+	+	+	-	-	-	+
22	Image color quantization with dither	-	-	-	-	-	+	+
23	Chromatic aberrations	+	-	-	-	+	+	+
24	Sparse sampling and reconstruction	-	-	-	-	+	-	+

One more distinctive feature of TID2013 is that there are five levels of distortions that approximately correspond to peak signal-to-noise ratio (PSNR) values equal to 33, 30, 27, 24, and 21 dB. This is usually enough for databases [13, 14] since these levels cover the most important range of distortions starting from almost invisible and ending by annoying ones. As the result, TID2013 contains 3000 distorted images (25 test images with 24 types and 5 levels of distortions).

Mean opinion score for each image was obtained as the result of experiments in which almost 1000 volunteers from 5 countries (Ukraine, Finland, Italy, France, and USA) took part. In each experiment, each volunteer was asked to compare 2 distorted images with the distortion-free etalon and to choose a better quality image. Tests were done separately for each reference image. Taking into account the recommended restrictions on test duration, each distorted image was participated in 9 comparisons where a winner was getting one point. Thus, MOS after averaging and removing abnormal judgments varies in the limits 0...9 where a larger MOS relates to better visual quality determined by subjects.

The fixed numbers of test images, distortion types and levels allow performing correlation analysis for all types (then, characterization of metric universality is obtained), for certain sets of distortions (where it becomes possible to analyze specific features of metrics like, e.g., sensitivity to color distortions, some groups are given in Table 1 and distortions included in them are marked by +) and for a particular type of distortion.

To characterize a relation between a quality metric and MOS, usually rank order correlation factors of Spearman (SROCC) and Kendall are employed since they do not require fitting operation (that can be done in different ways) needed to calculate standard Pearson correlation factor. The use of SROCC allows determining general accuracy without detailed analysis. A more detailed study can be performed if distortion types are collected in groups as it is shown in Table 1 where 7 groups are presented. The problem of evaluating quality for images with combined distortions is considered in the paper. Therefore, the group of multiple distortions ("MD") has been added to Table 1. There are the following MDs: #9 is filtering of noisy images, #21 is compression of noisy images, and #23 is a chromatic aberration (can be considered as a combination of blur and color component shifting).

Another database used by us in further analysis is the recently proposed LIVE MD. It contains distorted images for 15 test images, all of size 1280x720 pixels. There are 5 types of distortions where 3 are particular ones and 2 types of combined distortions of our interest, namely:

1) Blur;
2) JPEG;
3) Additive noise;
4) Blur followed by JPEG;
5) Blur followed by Noise.

Each particular distortion type (1...3) has 3 intensity levels with the chosen values of the corresponding parameters: parameter σ_G = 3.2, 3.9, 4.6 pixels, parameter of DCT matrix quantization Q = 27, 18, 12 and variance σ^2_N = 0.002, 0.008 and 0.032. These levels were chosen for perceptually separating of the resulting distorted images from each other and from the references. Meanwhile, these distortions were kept to be within a realistic range.

The combined distortions of types #4 and #5 are combinations of three levels of each already mentioned types of particular distortions with the same intensity levels. As the result, the LIVE MD consists of the two test sets (blur with additive noise and blur with JPEG) 225 test images each. There are 90 images with particular types of distortions (45 for each type) and 135 with the combined distortions.

Subjective assessment experiments have been done for each part separately and different volunteers have been attracted to them. Totally, 37 volunteers took part in experiments (19 and 18, respectively) who evaluated quality of each distorted image separately using 100-point scale (etalon images' quality has been assessed as well). Semantic labels 'Bad', 'Poor', 'Fair', 'Good' and 'Excellent' were marked at equal distances along the scale to guide the subjects. Visual quality has been finally determined as difference MOS, i.e., as difference between estimates for etalon and the corresponding distorted image.

For both considered databases and methods of image quality assessment for them, there are certain limitations. The database LIVE MD contains only two types of combined distortions – blur with noise and blur with JPEG and only three levels of distortions are considered.

A rather small number of volunteers have participated in experiments that influences verification accuracy. An example of estimation errors is demonstrated in Fig. 1 which shows MOS values vs. PSNR for blur that have been obtained for the same images for two independent subsets. Difference in MOS values can exceed 10, for example, for images 15b1 and 02b3 where index relates to the reference image and sub-index shows the type and level of distortion.

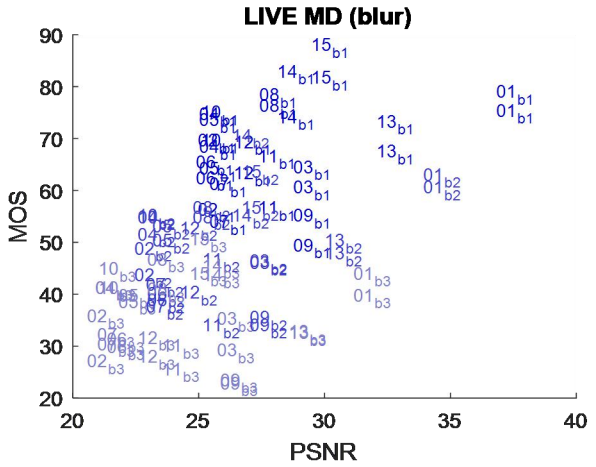

Figure 1. The scatter-plot of PSNR values for LIVE MD blur distortion vs. MOS

MOS accuracy is one of the key factors that determine the effectiveness of the metrics verification. To assess the impact of this factor and to determine the requirements to the number of necessary subjective experiments for the TID2013, these studies were carried out in [18]. A model of subjective experiment which took into account the following was constructed:

1) data of TID2008 experiments, which contain full details of each test pass;

2) the relative image quality estimation technique as described in detail above;

3) the model of subjective choice between two images at each stage of the comparison, depending on their quality, which also takes into account the probability of accidental clicking on one of them and subjective error of each image quality estimation.

As it was shown by the simulation results on TID2008, MOS accuracy comparing to the true quality values (subjective estimation error is equal to 0) at level of SROCC = 0.99 requires minimum 20 experiments for each test image and at least 50 to achieve the accuracy more than 0.995. It is necessary to consider the relative approach to estimation and using nine comparing steps in each experiment. Subjective experiments of TID2013 provide high accuracy, since about 1000 experiments, which give an average 40 for each reference image. The database LIVE MD has a sufficiently less accuracy since participants carried out less than 20 experiments for each image.

One drawback of images in TID2013 from the viewpoint of combined metric analysis is that the gradations of component distortions are not varied since their joint contribution is adjusted to fit five aforementioned PSNR values. Besides, distortion types in TID2013 have been chosen based on digital image processing applications. Taking into account the aforementioned limitations, below we employ data from both databases for increasing reliability of conclusions.

Verification of visual quality metrics

Performance analysis for multiple distortions is especially important for HVS-metrics which are the best for particular types of distortions. Because of this, we will consider metrics that have provided the best results for all types of distortions for TID2013. Recall that SROCC values exceeding 0.8 are provided by the following five metrics: FSIMc, SFF, PSNRHA [6] and PSNRHMA, SRSIM [19]. Some results are presented in Table 2. Recall that SRSIM operates with grayscale images or with luminance components of color images.

Table 2. Five the best metrics by the TID2013

#	Metric	Noise	Actual	Simple	Exotic	Color	MD	Full
1	FSIMc	0.902	0.915	0.947	0.841	0.775	0.94	0.851
2	SFF	0.879	0.906	0.95	0.821	0.832	0.904	0.851
3	PSNR-HA	0.923	0.938	0.953	0.825	0.632	0.869	0.819
4	PSNR-HMA	0.915	0.934	0.937	0.814	0.675	0.851	0.813
5	SRSIM	0.907	0.921	0.955	0.856	0.561	0.945	0.807

These results show that the analyzed metrics provide high SROCC values (about 0.9 and larger) for basic types of distortions met in practice (these types are collected in subsets “Noise”, “Actual” and “Simple”). Meanwhile, for the subset “Exotic” and, especially, the subset “Color”, the SROCC values are smaller and can be as low as 0.6...0.7. This means that these subsets contain particular distortion types that “cause problems” for the considered HVS-metrics. To determine them, Table 3 presents SROCC values for two types of distortions where additive white Gaussian noise (distortion type #1) is one of them.

Table 3. SROCC values for the most “problematic” distortion type in TID2013 (in pair with AWGN)

#	FSIMc	SFF	PSNR-HMA	PSNR-HA	SRSIM
1 & 2	0.913	0.841	0.889	0.904	0.920
1 & 6	0.791	0.843	0.861	0.869	0.806
1 & 12	0.870	0.838	0.776	0.816	0.865
1 & 14	0.810	**0.690**	**0.559**	**0.624**	0.764
1 & 15	0.721	0.747	0.789	0.799	0.841
1 & 16	0.876	0.830	0.854	0.882	0.847
1 & 17	**0.668**	0.708	0.765	0.798	**0.639**
1 & 18	**0.535**	0.840	**0.424**	**0.286**	**0.045**
1 & 23	0.894	0.806	0.696	0.759	0.912

The results presented in Table 3 allow determining problematic types of distortions for the considered metrics. For example, images with distortion type #14 (shifts of 8x8 pixel blocks with respect to their true position) are hard for three considered HVS-metrics (marked by bold). The reason for this is the following. Such shifts can be hardly noticed in textural test images and in homogeneous regions of test images although they clearly appear themselves at edges distorting their shape. If distortion has not been noticed, an observer might put a high mark

to image quality although a full-reference metric gets decreased value. Distortion type #17 deals with the contrast changing. The problem of dealing with this type of distortion is connected with the fact that people like contrast enhancement if it is natural (is not too large) while almost all metrics react to both contrast increasing and decreasing as to image degradation. Note that "favorable" contrast enhancements have not been determined yet and, thus, this peculiarity of HVS is not fully incorporated in existing HVS-metrics. Distortion type #18 that deals with color saturation appears only in color components. Four out of five considered metrics do not adequately assess visual quality of images with this type of distortions. This distortion type is the most problematic for the metrics PSNR-HA and PSNR-HMA. For the metric SRSIM such result could be expected since it does not react to color distortions at all. Distortion type #23 – chromatic aberrations is also rather hard for the metrics PSNRHMA and PSNRHA that have to be improved in the sense of taking into account color feature. The metric SFF has certain problems with this type of distortions too. Therefore, even the best HVS-metrics have problems with one or several types of distortions.

Scatter-plots for the metrics SFF and PSNR-HMA that demonstrate problems of assessing visual quality for images with distortion type #23 are presented in Fig. 2. Black color relates to points that correspond to images with distortion type #1 whilst blue color points relate to images with distortion type #23. Each point consists of several numbers where large digits relate to reference image index, the upper index is a distortion type and sub-index relates to the intensity level. In both cases, the metrics underestimate visual quality of images with distortion type #23 (MOS values are mostly larger for the same value of HVS-metric).

The method of analysis used above is based on correlation calculation for a pair of distortion types. It allows one to estimate not only an accuracy of ranking for a given type of distortion but also a degree of its coincidence for subjective judgments. The database LIVE MD has blur as the basic type of distortions. Thus, let us use this type as one of the distortion type in pairwise comparisons. For this purpose, let us calculate SROCC for images corrupted by blur and images with other types of distortions. The obtained results are presented in Table 4.

Table 4. Results of pair-wise rank correlation for blur and considered type

#	FSIMc	SFF	PSNRHMA	PSNRHA	SRSIM
1 & 2	0.903	0.907	0.821	0.789	0.908
1 & 3	0.856	0.856	0.786	0.749	0.860
1 & 4	0.749	0.768	0.520	0.468	0.758
1 & 5	0.789	0.794	0.577	0.538	0.787

The metrics FSIMc, SFF, and SRSIM that are all based on SSIM demonstrate similar results. The cases of combined distortions (types ## 4 and 5) are the most complicated for them although - SROCC values are about 0.75...0.8.

For the metrics PSNR-HMA and PSNR-HA, the results are worse although they have demonstrated quite good performance for TID2013 in aggregate. In general, the presented results demonstrate that it is worth improving performance for multiple types of distortions for all five considered metrics, especially for PSNR-HMA and PSNR-HA. Note that blur assessment errors can result in a biased estimation of the distorted image visual quality for images with multiple distortions that have blur as one of components (Table 4). Scatter-plots for each distortion for the metric PSNRHMA are presented in Fig. 3.

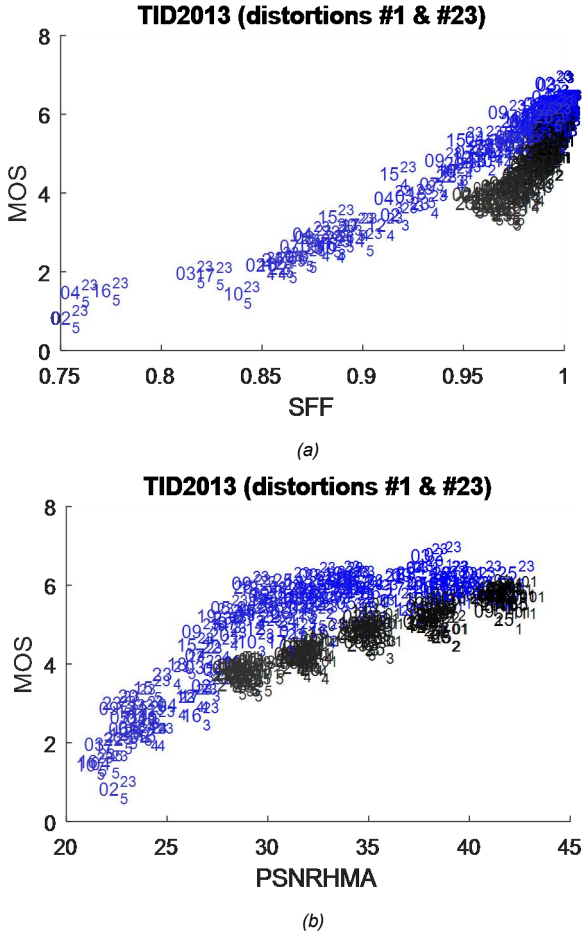

Figure 2. The scatter-plot of the metrics SFF (a) and PSNR-HMA (b) values for TID2013 distortions #1 and #23 vs. MOS

Analysis of scatter-plots shows that there is a sufficient dispersion of metric values for each distortion level. Let us consider some examples when, in our opinion, the metric PSNRHMA works erroneously. Fig. 3a presents the case of JPEG only distortions where we would like to mark the images 03j1 and 09j3 given in Figures 4a and 4b, respectively. MOS values for these images differ a lot (its values are equal to 86.6 and 56.5, respectively) while the metric has practically the same values - 35.8 and 36.2 dB, respectively. The image 09j3 has considerable blocking artifacts and less saturated colors. For the same type of distortion, the images marked as 12j3 and 09j1 have the same MOS=79 but the values of PSNRHMA values differ by 7.6 dB (they are equal to 31.9 and 39.4 dB, respectively). These data allow stressing the problem of PSNRHMA that consists in underestimation of quality for images with bright colors (#3 and #12). Besides, the metric has highly assessed the quality of the image 09j3 despite of rather high level of JPEG distortions.

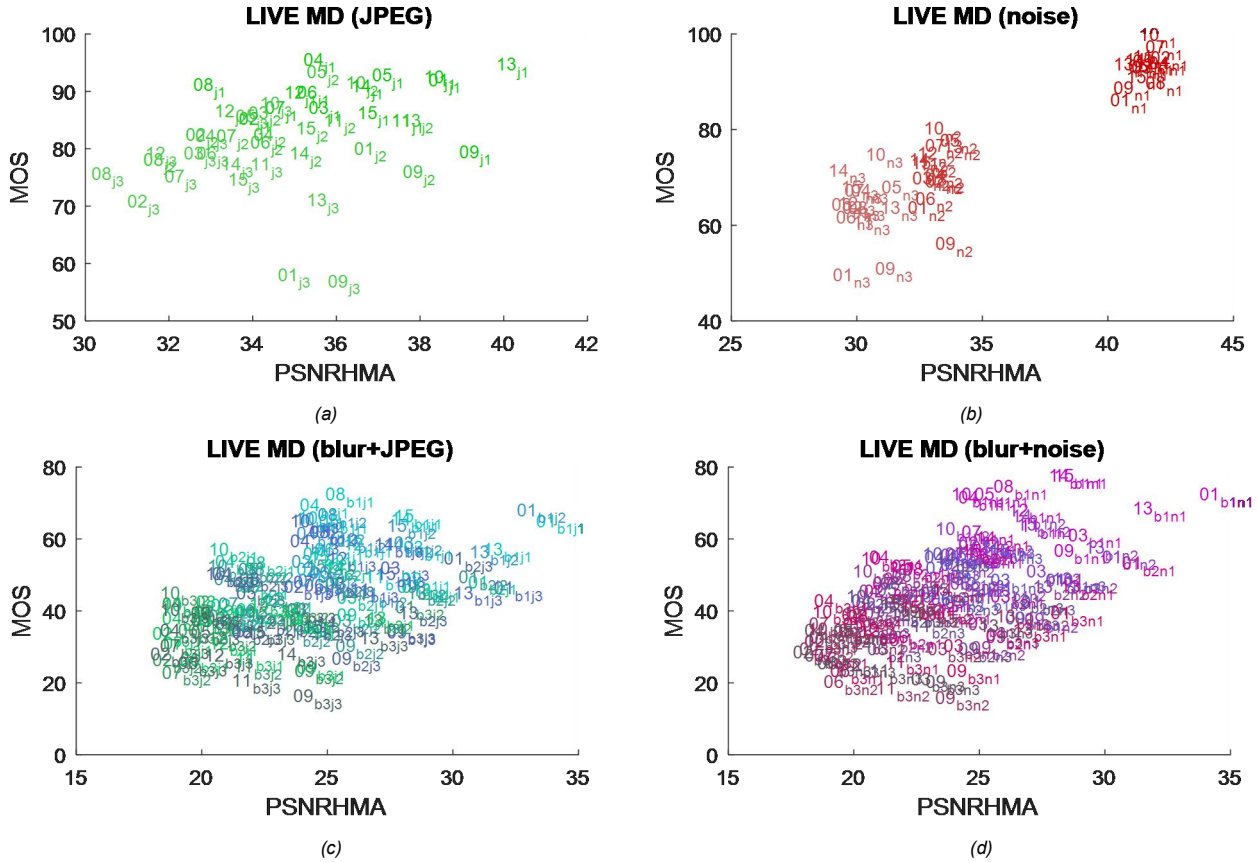

Figure 3. The scatter-plot of MOS vs. PSNR-HMA values for LIVE MD particular and combined distortions: JPEG (a), additive noise (b), blur+JPEG (c) and blur+noise (d)

Figure 4. MOS and PSNRHMA values for examples of single distortions: 03j1 (MOS = 86.6, PSNRHMA=35,8 dB) (a); 09j3 (MOS = 56.5, PSNRHMA=36,2dB) (b); 09n2 (MOS = 55.7, PSNRHMA = 33.9 dB) (c); 10n3 (MOS = 74.3, PSNRHMA = 31.1) (d).

Let us consider another example – the images 9n2 and 10n3 (see Figures 4c and 4d, respectively). They have the metric values equal to 33.9 dB and 31.1 dB whilst MOS values are equal to 55.7 and 74.3, respectively. In this case, the image 10n3 corrupted by the noise of higher intensity has, according to human judgments, better visual quality. Analyzing these images, we can note that at the image with palace there are large quasi-homogeneous image regions that correspond to fragments of sky and square. Noise in them is seen well whilst in fragments with details and textures that correspond to palace itself the noise is masked. The image #10 is the railway station from inside. Brick walls and peculiarities of architecture partly mask noise in almost all areas, probably, because of this is not noticeable well.

The key moment for the metric PSNRHMA is that it already takes into account masking effects of image fragments. However, this can be not perfect. Thus, we can expect that the metric performance can be improved by means of better incorporation of masking effects in the metric's calculation.

As it follows from the analysis of the results for the LIVE MD database, the test images there are "quite complex" for the metric PSNRHMA; even particular (not multiple) distortion types lead to considerable diversity of the metric values. Then, one can expect that increasing the number of distortion levels (nine for multiple distortions instead of three for particular distortion types) will lead to higher diversity of estimates (see data in Figures 3c and 3d) and, in turn, to smaller SROCC.

Analysis of the scatter-plot confirms some conclusions given above. Images with essential contribution of JPEG distortions have overestimated quality (for the image 01b1j3 the metric value exceeds 32 dB) whilst larger MOS values are observed for images with predominant blur contribution. Figures 5a and 5b present images with multiple distortions 01b3j3 and 04b1j1 (blur+JPEG) for which the metric and MOS values disagree to each other (MOS = 34 and PSNRHMA = 28.4 dB vs MOS = 69 and PSNRHMA = 25 dB, respectively).

Similar situations are observed for multiple distortions of the type noise+blur. The corresponding scatter-plot in Fig. 3d is characterized by smaller diversity of metric values and larger compactness of data for the considered distortion levels. This means that the metric performs more adequately for this situation. The problem with incorporating masking effect for the case of these distortions is less than for the already considered cases. One possible reason is that in forming distorted images noise is added to already blurred images. Due to this, details are smeared and their masking effect decreases.

Despite of the aforementioned drawbacks, the databases LIVE MD and TID2013 allow detecting disadvantages and problematic situations even for the best HVS metrics that, for other simpler databases, might have SROCC values of about 0.9…0.95, i.e. to seem almost perfect.

Using the metric PSNRHMA as an example, we have demonstrated how to determine the reasons why this metric does not provide SROCC close to 0.9 or larger for several types of distortions in TID2013 and LIVE MD. In particular, it has been shown that the main problem is with color distortions that have been confirmed by data for distortion type #18 in TID2013. It is also desirable to improve the modeling of masking effects by better incorporation of peculiarities of HVS.

(a)

(b)

Figure 5. MOS and PSNRHMA values for examples of multiple distortions: 01b3j3 (MOS=33.9, PSNRHMA=28.4 dB) (a); 04b1j1 (MOS=69, PSNRHMA=25 dB) (b).

Modifications of PSNRHMA

To improve adequacy of PSNRHMA, we have carried out several modifications with the main intention to improve performance for combined distortions by the TID2013. The following changes have been introduced and analyzed:

1) More accurate values of contrast sensitivity function (CSF) for DCT blocks of size 8x8 pixels in color space YCbCr have been used. In the original metric, these values have been calculated based on the quantization table recommended for JPEG standard [20]. Because of some limitations of this table, researchers have proposed many modifications for JPEG image processing [21, 22]. According to verification on TID2013, one of tables proposed in [21-23] was chosen.

2) The metric PSNR-HMA contained an algorithm of contrast change and accounting a mean shift. Optimization of it was carried out for TID2008. New distortion types were added to TID2013, and most of them are relate to color components. Therefore, considering this fact, especially low SROCC for distortion #18 (change of color saturation), weight values can be redefined separately for luminance and color components.

3) The paper [24] presents description of calculating the masking effect. To exclude overestimation of its influence in heterogeneous blocks of size 8x8 pixels (edges in the first order), it was proposed to calculate the correcting factor (δ in the equation (2) in [24]) as

$$\delta = \frac{V_1 + V_2 + V_3 + V_4}{4 \cdot V_{sum}}, \quad (1)$$

where V is a local energy of a block of size 8x8 pixels and $V_1 \ldots V_4$ denote local energies of 4x4 pixels blocks that comprise the given block. The main problem of this approach describing the masking effects arises if edge is not straight and has an orientation close to diagonal. The second modification deals with more universal and accurate detection of heterogeneities in 8x8 pixel blocks. We analyze below several methods of edge detection and gradient mapping such as Sobel, Canny, Roberts, Prewitt, logarithmic. Good performance has been recently demonstrated by phase congruency employed in the metric FSIMc [4]. Thus, this detector has been used in analysis as well. Below we present only the best results. Note also that edge detector performance depends upon a selected threshold.

4) It is also desirable to take into account a heterogeneity area. For this purpose, we propose to use the following formula for the correcting factor:

$$\delta = C \cdot \frac{\sum_{i=1}^{8}\sum_{j=1}^{8} I_{ij}^{P}}{64}, \qquad (2)$$

where I_{ij} is an 'informativity' (characterized by the presence of non-uniform image areas with actual information) value for the pixel with indexes *i* and *j*. For calculating I_{ij} we used the methods that were previously discussed in item 3). All of them have been normalized, with their values varying in the range 0…1.

Since the methods have different sensitivity to heterogeneities, informativity values of a single pixel can differ significantly and affect the masking effect of the block. Taking this into account and examining its impact, the coefficient P with values (1/3, 1/2, 1, 2 and 3) has been added.

The coefficient C is added to account for nonuniformity in a block of 8x8 pixels is determined by the following expression:

$$C = \frac{1}{64 - N(I_{ij} > T) + 1}. \qquad (3)$$

The parameter $N(I_{ij} > T)$ was added due to the following reason. In the case of quasi-homogeneous area values of I_{ij} are close to 0, therefore the masking effect for an image block of 8x8 pixels is almost absent. For a non-homogeneous area occupying the entire block, we have an opposite result. Its values of I_{ij} are higher and depend on the texture characteristics. But on an edge with a significant difference of luminance value, an informativity for the corresponding pixel is also close to 1, whereas the block has no such masking properties. Therefore, the correction parameter has been added which determines the number of pixels in the block with values of I_{ij} higher than threshold T (see Fig. 6).

As we can see, for the texture occupies the entire block, parameter C is equal 1 and does not affect on the value of the masking effect. For partially uniform block we reduce it.

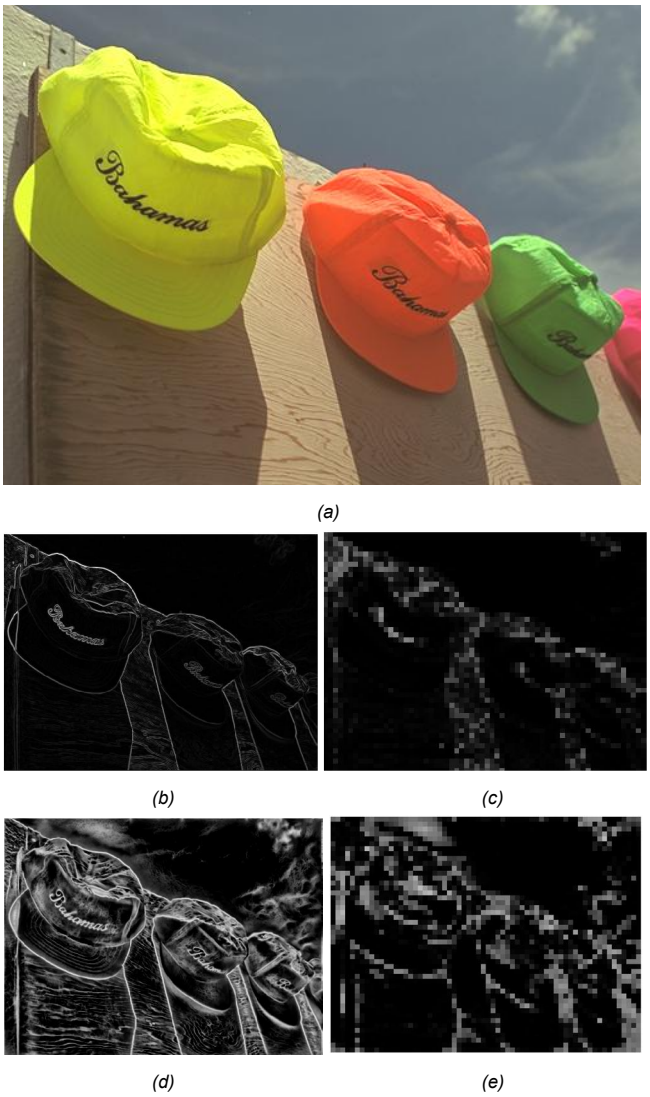

Figure 6. Test mage (a), its gradient map (b) with δ (c) and phase congruency (d) with δ (e)

Using (2), the study of modification performance and expedience of their use has been done. Some of the results are presented in Table 5 for different methods of calculation of block informativity, different functions I_{ij}^{P} and thresholds. The introduced modification of CSF has been employed in all six variants of the modified metric given in Table 5, the sign "–" mean that the corresponding modification has not been used. In the column "Informativity", the following notations are used for the studied detectors: 1 – Canny, 2 – phase congruency, 3 – gradient (Prewitt). The column Threshold presents the threshold values used in (2).

As it follows from the analysis of data in Table 5, the modifications lead to metric performance improvement. More accurate values of CSF using new quantization table result in improved results for all distortions (see the rightmost column "Full") as well as for subsets "Color". The modified informativity, in general, provides better results as well although the influence of the used informativity parameter and threshold can be essential.

Table 5. Results for the best modifications of the PSNR-HMA

#	Informativity	Threshold	Noise	Actual	Simple	Exotic	Color	MD	Full
PSNR-HMA			0.915	0.934	0.937	0.814	0.675	0.85	0.813
1	-	-	0.911	0.931	0.933	0.8	0.73	0.828	0.821
2	**1**	-	0.913	0.934	0.934	0.81	0.722	0.835	0.827
3	**1**	**0.2**	0.919	0.938	0.946	0.812	0.71	0.841	0.83
4	**2**	**0.2**	0.915	0.933	0.948	0.813	0.72	0.848	0.825
5	**3**	**0.2**	0.92	0.938	0.949	0.813	0.7	0.846	0.831
6	**3**	**0.1**	**0.922**	**0.939**	**0.951**	**0.828**	**0.805**	**0.862**	**0.854**

We have considered several methods of informative region selection with most quality work of two of them: Prewitt and phase congruency. Their examples are shown in Fig. 6b and 6d. However, in the proposed FSIMc, phase method has several disadvantages. It is too sensitive to gradients and edges, thereby underestimating the texture masking effect. The second problem with it is a large computational complexity. The computation of the modification will take more than 4 times longer (more than 1 second for an image size of 512x384). Therefore, we chose the Prewitt method.

The last step of the new metric optimization is a recalculation of the contrast change and mean shift. Mean shift coefficient is remained almost unchanged (it has changed from 0.04 to 0.045). Original contrast weights of 0.002 and 0.25 for new Y component were changed to 0 and 0.37. For color components, new values became 0.25 and 2, respectively. Final verification of the metric on TID2013 allows determining optimal values of threshold (0.1) and power P = 2 in (2).

The obtained result outperforms all considered metrics for groups “Actual” and “Full”. All changes (see the results for the modification #6) positively impact the metric’s performance for all groups, especially for “Full” and “Color”. For the group of combined distortion, SROCC of modification #6 has increased by 0.01. In Table 6, the results for the image database LIVE MD are presented.

Table 6. Results of the 6th modification for LIVE MD

Metric	Pairs of distortion			
	1 & 2	1 & 3	1 & 4	1 & 5
PSNRHMA	0.821	0.786	0.520	0.577
Mod #6	0.795	0.753	0.484	0.550

Although Accuracy has increased for groups "Noise", "Actual" and “MD” by the 0.01-0.02, values for pairs of combined distortions in LIVE MD (with noise and JPEG compression distortion) have decreased, as it includes a noise and distortion JPEG. We hope that better ways to calculate correcting factor can be found in future to provide an appropriate trade-off. Alternatively, weights for intensity and color components using different CSFs can be applied. Perhaps, effects of regions of interest, macro and portrait imaging can be somehow taken into account as well.

Conclusions

The paper presents result of performance analysis of full-reference metrics of visual quality for images corrupted by multiple distortions. Verification of several metrics that belong to the best known ones (FSIMc, SFF, PSNRHMA) has been carried out for the databases TID2013 and LIVE MD using specific methodology of analysis. This has allowed us to determine the main issues (groups of “unfavorable” distortion types) for the existing metrics and the methods of their improving.

It has been shown that one of the main aspects is an adequacy to color distortions. One reason is that many existing metrics have been designed for grayscale images and later they have been modified and applied to color images.

An important way to improve metric performance is to better model HVS sensitivity to color distortions and masking effects. Other peculiarities of HVS such as regions of interest are worth considering too. As a particular case, modifications are introduced to the metric PSNRHMA to partly correct drawbacks detected for it. Due to these modifications, SROCC has been increased by more than 0.04 for the database TID2013, reaching 0.854. This was due to better correspondence for new types of distortions in this database, especially, color ones.

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] 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.

[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] 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.

[8] 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.

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

[10] 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.

[11] 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. 2016.

[12] Y. Horita, K. Shibata, Z.M. Parvez Saddad, Subjective quality assessment toyama database, in http://mict.eng.u-toyama.ac.jp/mict/, 27 Nov. 2016.

[13] 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.okstate.edu/?loc=csiq, 7 Dec. 2016.

[14] 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, 27 Nov. 2016.

[15] P. Le Callet and F. Autrusseau, “Subjective quality assessment IRCCyN/IVC database,” 2005. 2012. [Online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/[Accessed 25 Nov. 2016].

[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. 2016].

[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] O. Ieremeiev, N. Ponomarenko,V. Lukin, "Evaluation of accuracy characteristics for the verification techniques of visual quality metrics based on specialized image databases [in Russian]," Radioelectronic and computer systems, vol. 2(61), pp. 48-57, 2013.

[19] L. Zhang and H. Li, “SR-SIM: A fast and high performance IQA index based on spectral residual,” in Proc. of 19th IEEE Int. Conf. on Image Processing (ICIP), 2012, pp. 1473 - 1476.

[20] W. B. Pennebaker, J. L. Mitchell, "JPEG Still Image Data Compression Standard " – Norwell: Kluwer Academic Publishers, 1992. – 638 p.

[21] J. Chao, H. Chen and E. Steinbach “On the design of a novel JPEG quantization table for improved feature detection performance,” in Proc. of 20th IEEE Int. Conf. on Image Processing (ICIP), 2013

[22] M. Konrad, A. Uhl, "Evolutionary Optimization of JPEG Quantization Tables for Compressing Iris Polar Images in Iris Recognition Systems," in Proc. of 6th Int. Symp. on Image and Signal Processing and Analysis (ISPA), 2009, pp. 534-553.

[23] JPEG Compression Quality tables for Digital Cameras and Digital Photography Software [Online]. Available: http://www.impulseadventure.com/photo/jpeg-quantization.html [Accessed 7 Dec. 2016].

[24] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, “On between-coefficient contrast masking of DCT basis functions,” in Proc. of the 3rd Int. Workshop on Video Processing and Quality Metrics, 2007, 4 p.

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.