IMAGE VISUAL QUALITY METRICS VERIFICATION BY TID2013: EXPLORING OF MEAN SQUARE ERROR DRAWBACKS

Nikolay Ponomarenko ([1]), Vladimir Lukin([1]), Oleg Ieremeiev ([1]), Benoit Vozel([2]), Kacem Chehdi([2]), Karen Egiazarian ([3]), Jaakko Astola ([3])

([1]) National Aerospace University, Kharkov, Ukraine
([2]) University of Rennes 1, Lannion, France
([3]) Tampere University of Technology, Tampere, Finland

ABSTRACT

Specialized image databases like TID2013, allow quantitative evaluation of adequateness to human perception characterized by mean opinion score (MOS) and a verified full-reference visual quality metric using certain criterions such as rank order correlation coefficients (ROCC). In this paper, we propose modifications of known Spearman and Kendall ROCCs that take into account the fact that MOS for existing databases has been measured with a limited accuracy and is most certainly erroneous (contains "noise component"). For the database TID2013 we have also marked images with practically invisible distortions. This has led to considering criteria for detecting such images. Using a mean square error we also analyze metrics based on alternative norms l^1, l^3, and l^4. Analysis is performed for the database TID2013 and different sensitivity of human vision system (HVS) to distortions in color and intensity is taken into account. One approach to analyze adequateness of metrics for different types of distortions is described and proved exploitable.

Index Terms— Image quality, image color analysis, detection algorithms, correlation coefficient, mean square error methods

1. INTRODUCTION

Full-reference image visual quality metrics [1] are widely used in various applications of image and video processing. They, in particular, include lossy compression, image/video denoising, digital watermarking, etc. Efficiency of solving these tasks considerably depends upon an adequateness of used HVS-metric to image/video perception by humans. For example, if an inadequate metric has been exploited in a design of image denoising method, then it is hard to expect that a designed method would occur efficient in practice.

It is difficult to present an adequateness of a given metric to human perception by some mathematical expression. Because of this, researchers employ several standard statistical approaches applicable for such large databases as TID2008 [2] or LIVE [3]. A large number of volunteers (observers, participants) are attracted to subjective experiments intended for obtaining mean opinion score for each image in a database. Then, all images of a given database are used to verify a given metric for the obtained MOS array by calculating correlation factors. Rank order correlations such as Spearman (SROCC) and Kendall (KROCC) [4] are used more often to avoid data overfitting. Their larger (approaching to unity) values show that there is a good correspondence between an analyzed metric and image perception (visual quality assessment) by humans.

Values of MOS for any image database are based on opinions of individual observers that are subjective and can be influenced by such errors as inattention, randomly pressing a wrong button (grade, image) and so on. Besides, a number of experiments and comparisons for each particular image is limited in spite of huge efforts and time spent on performing experiments. However, note that alongside with MOS array offered for each database, the database creators often present the values of standard deviation (STD) or variance for each MOS.

This STD values characterize accuracy of experimental data. However, the observed errors in MOS are not taken into consideration in calculating SROCC and/or KROCC and in metric verification. To partly get around this shortcoming, below we introduce modified versions of SROCC and KROCC that are able to incorporate data on STD of MOS for images. These modified SROCC and KROCC are later used in our analysis.

A statistical estimate of metrics correspondence to a human perception is better if a database is more representative, i.e. contains more distorted images. The largest openly available database is TID2013 [5] that contains 25 test images with 24 types of distortions and 5 levels of distortions. Therefore, in aggregate, the database includes 3000 distorted color images. However, such a large number of images forces to carry out experiments for MOS obtaining with dividing distorted images into subsets. For example, for TID2008 and TID2013, the subjective experiments have been carried out separately for subsets of distorted images that relate to each of 25 etalon images. To get around this problem, we modify methodology of ROCC correlation that takes into account aforementioned peculiarity to obtain MOS.

A typical image processing task dealing with image visual quality is to provide an invisibility of introduced distortions. Examples are lossy image compression and digital watermarking. A good metric should be able to detect a visibility of introduced distortions. It is desirable to analyze this at the stage of metrics verification. We show how this appears possible due to marking images with not noticeable (invisible) distortions in TID2013.

It has been shown in [5] that a good HVS metrics should take into account color information and that for certain distortion types there are underestimation or overestimation of visual quality. Using the database TID2013, we analyze these factors on the standard metric Mean Square Error (MSE) as well as on the metrics based on other norms, l^1, l^3, and l^4.

2. PROPOSED MODIFICATIONS TO A VERIFICATION PROCESS

As it was mentioned, useful information on metric adequateness is provided by SROCC and/or KROCC that characterize difference in positions of images in sorted array of a metric and sorted array of MOS values. However, due to aforementioned errors in estimating MOS, there are situations when not only a considered metric is not ideal but also MOS is “noisy” and, thus, its position in the sorted array is displaced (with respect to a position in case of number of experiments approaching infinity). For example, MOS of the image “I01_01_1.bmp” in TID2013 is equal to 5.51 whilst for the image “I01_01_2.bmp” one has MOS 5.57. Both images are corrupted by additive white Gaussian noise where for the latter image the noise variance is larger. Then, visual quality of the former image has to be higher and its MOS should be larger but this does not happen. However, taking into account that MOS STD values for these images are equal to 0.13 and 0.16, respectively, the observed difference in MOS values can be expected. However, such a fact leads to decreasing ROCCs values.

To decrease the influence of such errors in MOS on calculated ROCC, let us introduce the modified KROCC K^R calculated as

$$K^R = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \delta(i, j) , \qquad (1)$$

where n is the number of estimates of MOS, $\delta(i, j) = \begin{cases} 1, & M_i - 2\sigma_i \le M_j \\ -1, & M_i - 2\sigma_i > M_j \end{cases}$, M_i is a value of i-th element of the MOS array arranged in ascending order according to metric values, σ_i denotes the corresponding STD.

Expression for the modified SROCC can be written as:

$$S^R = 1 - \frac{6}{n(n+1)(n-1)} \sum_{i=1}^{n} \lambda(i)^2 , \qquad (2)$$

where $\lambda(i)$ denotes a difference between ranks of i-th image in sorted (in ascending order) arrays of a considered metric and MOS. Here, in a difference calculation, the images for which their MOS differs from MOS of i-th image by no more than $2\sigma_i$ are not taken into account. If σ=0, expressions (1) and (2) reduce to standard KROCC and SROCC.

Since MOS values in TID2013 have been formed separately for each of 25 test (etalon) images [5], modified SROCC and KROCC can be separately determined for each subset of 120 images and then averaged

$$K^{int} = \frac{1}{25} \sum_{l=1}^{25} K_l^R , \; S^{int} = \frac{1}{25} \sum_{l=1}^{25} S_l^R \qquad (3)$$

where K^R_l and S^R_l are modified KROCC and SROCC calculated according to (1) and (2) for distorted images relating to l-th etalon image.

Table 1 contains data for several metrics. We present conventional values of SROCC and KROCC given in [5] and the modified values determined by (3).

Table 1. Values of conventional SROCC and KROCC and the values of S^R, K^R, K^{int} and S^{int} for several metrics

Metric	SROCC	KROCC	S^R	K^R	S^{int}	K^{int}
FSIMc [6]	0.85	0.67	0.92	0.78	0.93	0.81
PSNR-HA [7]	0.82	0.64	0.90	0.76	0.90	0.77
MSSIM [8]	0.79	0.61	0.87	0.72	0.89	0.75
SSIM [9]	0.64	0.46	0.76	0.58	0.78	0.60

As it follows from data in Table 1, the metric FSIMc [6] remains to be the best according to the modified ROCCs and SSIM [9] remains the worst. Meanwhile, K^{int} is always smaller than the corresponding S^{int} similarly as a conventional KROCC is smaller than the corresponding SROCC.

Note that K^{int} is larger than the corresponding KROCC and S^{int} is larger than the corresponding SROCC. This is because the influence of errors present in the obtained MOS has been reduced. Note that advantages of the metrics FSIMc and PSNR-HA can be attributed to the fact that they take into account color information [5].

Let us consider now an invisibility of distortions and detection of such images by HVS-metrics. To analyze this, all distorted images in TID2013 have been visualized and viewed by a group of experts to determine and mark those images for which distortions are practically (with high probability) invisible. Totally 214 out of total 3000 images in TID2013 have been marked.

As it can be expected, visibility of distortions depends upon image content and distortion type. Fig. 1 shows the number of distorted images for each group relating to each etalon image (25 totally) and Fig. 2 illustrates a dependence on distortion type (24 totally).

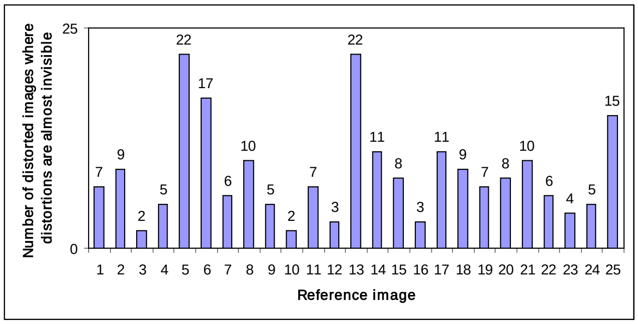


Fig. 1. Number of images with invisible distortions for each reference (etalon) image

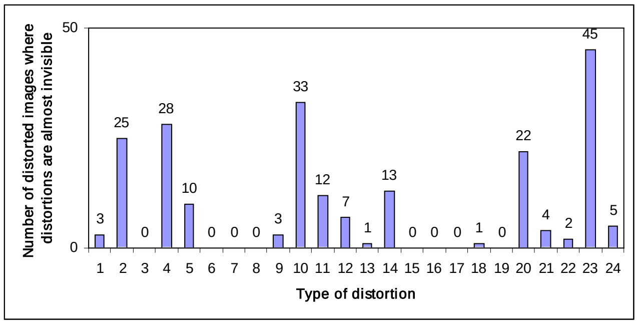


Fig. 2. Number of images with invisible distortions for each type of distortion

As it can be seen in Fig. 1, invisible distortions more often take place for reference images which either contain a lot of fine details or are highly textural (images #5 and #13, respectively). Concerning distortion types (Fig. 2), the "most invisible" type in TID2013 is a distortion type #24 (chromatic aberrations). Among "leaders" in this sense, there are distortion types #10 (due to JPEG compression), #4 (masked noise), #2 (additive noise in color components), #20 (comfort noise). These results one more time support known observations on different sensitivity of HVS to distortions in color and intensity as well as an influence of masking effects.

Availability of the marked images with invisible distortions allows analyzing performance of different metrics for detecting such images. Usually, detection is characterized by, at least, two probabilities. However, recently a new parameter called AUC [10] that provides an opportunity to characterize a detection by only one value in the limit from 0 to 1 (perfect detection) has been introduced:

$$A=(S_0-n_0(n_0+1)/2)/n_0/n_1, \qquad (4)$$

where n_0 and n_1 are the numbers of positive and negative examples, S_0 is sum of ranks of positive examples in the ranked list.

Table 2 presents the values of AUC for detecting images with invisible distortions in TID2013 using different HVS-metrics. As it can be seen, FSIMc and MSSIM are the best whilst PSNR-HA and SSIM produce worse results.

Table 2. Values of AUC for some HVS-metrics

FSIMc	PSNR-HA	MSSIM	SSIM
0.90	0.84	0.88	0.81

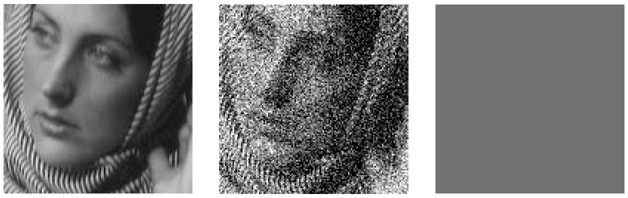
3. ANALYSIS OF MEAN SQUARE ERROR

MSE and peak signal to noise ratio (PSNR) which are conventional metrics in image processing have been many times reported as not adequate for characterizing image visual quality. However, the operation of MSE estimation in several modified forms is still present in many visual quality metrics [7]. Thus, it is worth starting to consider the ways to improve a metric performance just from MSE and its closest "derivatives". Recall that MSE is calculated as

$$\mathrm{MSE}=\frac{1}{\mathrm{WT}}\sum_{i=1}^{W}\sum_{j=1}^{T}(A_{ij}-A_{ij}^{n})^2 , \qquad (5)$$

where A is reference image, A^n is an image with distortions, i and j are pixel indices, W and T define image size.

Let us present an extreme example showing drawbacks and inadequateness of MSE. Image in Fig. 3,b is corrupted by AWGN with variance 2229. The image in Fig. 3.c is the homogeneous image that has the same mean as the test image Barbara in Fig. 3,a and its MSE is also equal to 2229.



a) b) c)

Fig.3. Example of inadequateness of MSE to human perception: a) reference image, b) image corrupted by AWGN with variance 2229, c) homogeneous image that also has MSE=2229 with respect to the reference image

As it is seen, metrics that use pixel-by-pixel image comparison are often unable to adequately characterize image visual quality. Other examples are possible. However, our study is intended on searching and analyzing other tendencies using the database TID2013. In particular, alongside with MSE that exploits l^2 norm, it is possible to use other alternatives as mean absolute error (MAE) based on l^1 as well as the metrics E^{l3} and E^{l4} based on the l^3 and l^4 norms:

$$E^{l3} = \frac{1}{WT}\sum_{i=1}^{W}\sum_{j=1}^{T}\left|A_{ij}-A_{ij}^{n}\right|^{3},$$

$$E^{l4} = \frac{1}{WT}\sum_{i=1}^{W}\sum_{j=1}^{T}(A_{ij}-A_{ij}^{n})^{4} . \qquad (5)$$

Table 3 presents the values of S^{int} and K^{int} for these metrics calculated and averaged for the three components of color images in RGB color space. MSE has the largest correlation but the results for the metrics E^{l3} and E^{l4} are quite close and better than those for MAE.

Table 3. Values of S^{int} and K^{int} for the considered metrics

Criterion	Metric			
	MAE	MSE	E^{l3}	E^{l4}
S^{int}	0.65	0.81	0.80	0.77
K^{int}	0.48	0.63	0.61	0.58

Note that while calculating S^{int} and K^{int}, the values of these metrics have been taken with the opposite (negative) sign to provide increase of the transformed metric values for higher visual quality (and MOS) of images in TID2013.

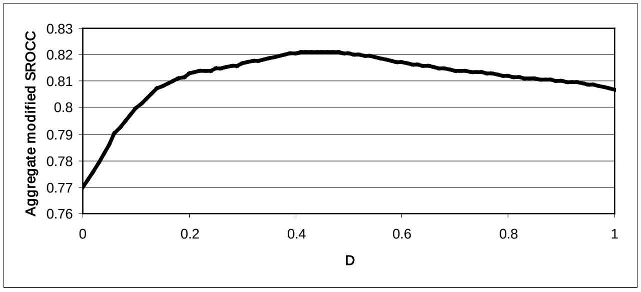


Fig. 5. Dependence of S^{int} on D for the metric MSE

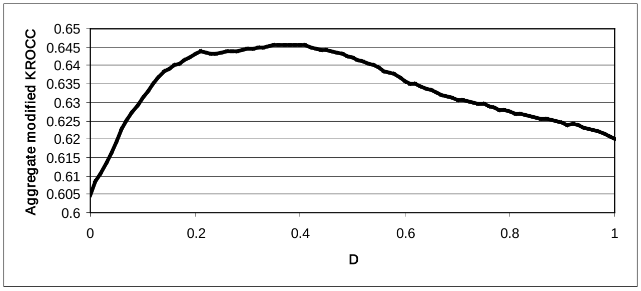


Fig. 6. Dependence of K^{int} on D for the metric MSE

Although many metrics that do not take into account color information are calculated for R, G, and B components separately and then averaged, this way is not the best [5]. The database TID2013 contains many images with specific color distortions and this allows taking into account different sensitivity to distortions in color and intensity components using YCbCr color space. Suppose that a metric is calculated as a weighted sum of this metric values for the intensity component Y (this weight is fixed and equals to unity) and color components Cb and Cr (for both components we used the same weight D). Figures 5 and 6 show the dependences of the modified SROCC and KROCC on D for the metric MSE.

Table 4 presents approximately optimal D for the considered metrics as well as the corresponding values of S^{int}, K^{int}, and AUC. As it is seen, after weight optimization, the metric MSE remains to be the best according to all three analyzed criteria. The results of analysis show that it is worth designing HVS-metrics based on the norm l^2. Thus, we continue our analysis for only MSE. Let us denote the weighted MSE as MSEw.

Table 4. Results of optimizing D for the considered metrics

Criterion	Metric			
	MAE	MSE	E^{l3}	E^{l4}
D	0.38	0.41	0.52	0.67
S^{int}	0.65	0.82	0.81	0.79
K^{int}	0.48	0.65	0.62	0.60
AUC	0.79	0.86	0.83	0.81

Figures 7 and 8 present the values of the modified SROCC and KROCC for each reference image of TID2013.

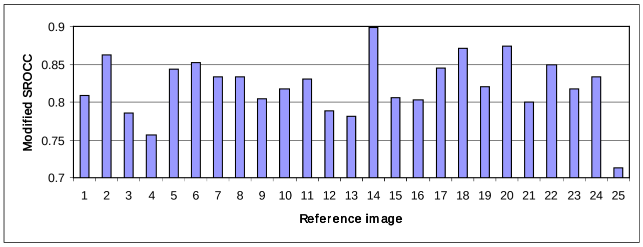


Fig. 7. Modified SROCC between MSEw and MOS for each reference image of TID2013

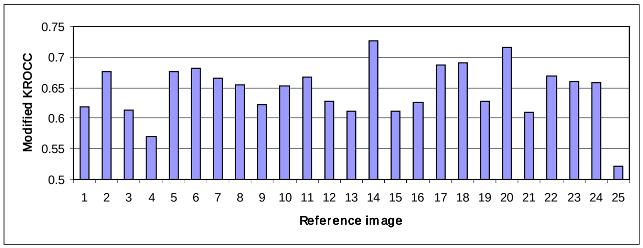


Fig. 8. Modified KROCC between MSEw and MOS for each reference image subset of TID2013

Diagrams in Figures 7 and 8 show that the reference images #25, #4, and #13 are the most complex for adequate characterization by MSEw. The images #13 and #4 are textural and MSEw does not take masking effects typical for these images. Complexity of the reference image #25 consists in the fact that it is artificial and contains different contrast noise-like texture whilst homogeneous areas are absent.

Fig. 9 presents the plots of first and second type errors in detecting images for which distortions are not visible. Detection has to be done by setting a certain threshold of PSNRw (or, equivalently MSEw). Here we present data for PSNRw connected with MSEw as PSNRw = $10\log_{10}(255^2/\text{MSEw})$.

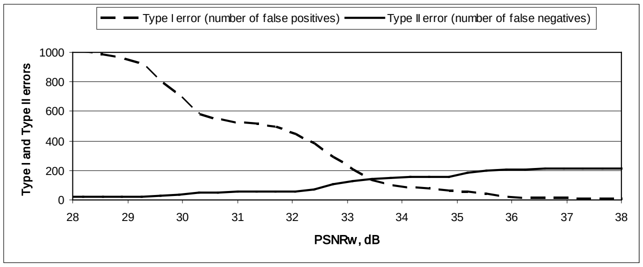


Fig. 9. Dependences of the first and second type error numbers on PSNRw for detecting images with invisible distortions in TID2013

As it is seen from an analysis of the dependences in Fig. 9, PSNR (and, respectively, MSE) is unable to serve well for detecting images with invisible distortions. Suppose that we need 90% of such images to be detected (i.e., 193 out of 214 images with invisible distortions) and this is achieved for the threshold PSNR_t=28 dB. However, the number of false positives (i.e., wrongly detected images) is over 1000 and exceeds the number of true positives by about 5 times. To diminish the number of falsely detected images with invisible distortions (e.g., to provide not more than 5% of false detections), one has to set PSNR_t=33.5 dB. However, in this case the number of false positives is more than half (132 false positives, 70 true positives). Thus, a question what threshold to set does not have a perfect answer and the setting is problematic.

Let us consider now what types of distortions are "most unfavorable" for MSEw. For this purpose, we have modified an expression (2) to determine a contribution β(j) of each distortion type #j in TID2013 to the sum $\lambda(i)^2$:

$$\beta(j) = 100\% \sum_{i=1}^{n} \vartheta(i)\lambda(i)^2 \Big/ \sum_{i=1}^{n} \lambda(i)^2 , \qquad (5)$$

where $\vartheta(i)$ equals to 1, if an i-th image relates to distortion type #j and to 0 otherwise.

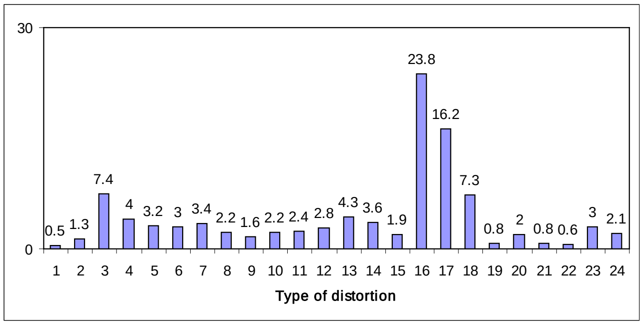


Fig. 10. Contribution into decrease of the modified SROCC of distortion type in TID2013

As it can be seen, additive white Gaussian noise (distortion type #1) has the smallest contribution to reduction of SROCC and, therefore, it is the simplest distortion type of MSEw (and this feature of MSE is widely exploited). Let us use the data for this distortion factor as reference curves in our further analysis. More in detail, we will analyze dependences of MOS on MSEw for different other types of distortions. If a curve for a given distortion type goes "below" the reference curve, then the metric "decreases" distortion level (overestimates the quality) and vice versa [5].

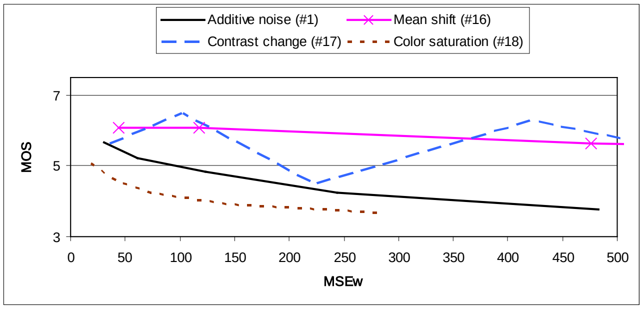


Fig 11. MSEw vs MOS for distortion types #1,#16,#17,#18

Fig. 11 presents dependences for distortions due to mean shift change (#16) and contrast change (#17) which are the most complex for MSEw (see data in Fig. 10) and color saturation distortions (#18). As it is seen in Fig. 11), MSEw considerably overestimates distortions for mean shift change and contrast change for the cases of enhanced contrast (that correspond to MSEw about 100 and 430). Meanwhile, for color saturation the quality of the corresponding images is overestimated.

Fig. 12 presents the plots for distortion types that relate to image denoising tasks.

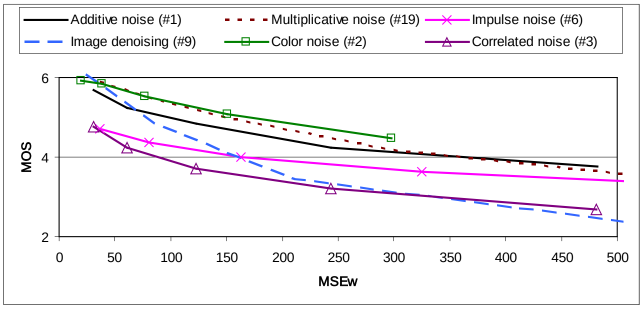


Fig 12. MSEw vs MOS for distortion types #1, #2, #3, #6, #9 and #19 relating to image denoising task

MSEw overestimates quality for impulse noise distortions as well as for residual distortions after denoising or for spatially correlated noise. Meanwhile, level of distortions of the type #2 (noise in color components) is still overestimated by MSEw.

This drawback could be slightly compensated by using D smaller than recommended (see Table 4). However, this leads to additional overestimation of image quality for the distortions of type #18 (see data in Fig. 11).

Fig. 13 presents dependences for distortion types #1, #4, #10, #11, #23, and #24.

MSEw overestimates quality for JPEG (#10) for small compression ratios (that correspond to small MSEw) and underestimates quality for high CR.

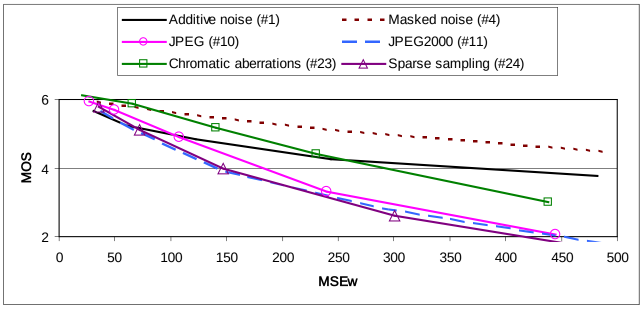


Fig 13. MSEw vs MOS for distortion types #1, #4, #10, #11, #23 and #24

Similar properties are observed for chromatic aberrations (#24) (for small distortion level they are masked and these effects are not taken into account by MSEw). For JPEG2000 (#11) and sparse coding (#25), image quality is overestimated for all distortion levels. For the masked noise (#4), MSEw overestimates distortion level since it does not take into consideration masking effects.

Fig. 14 presents the results for distortion types #1, #5, #7, #8, #14, and #20.

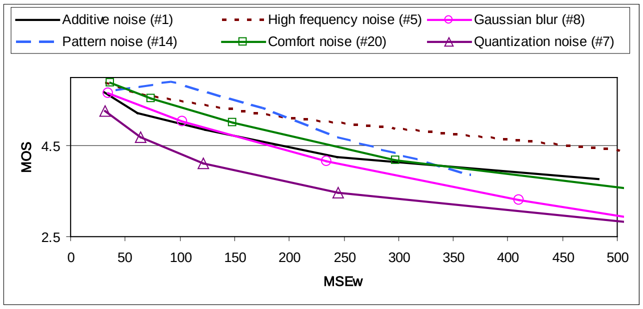


Fig 14. MSE vs MOS for distortion types #1, #5, #7, #8, #14 and #20

As it is seen, MSEw underestimates distortion level for quantization noise (#7) and high levels of blur (# 8). Meanwhile, this metric considerably overestimates distortion level for high-frequency noise (#5). This is because this metric is unable to take into account different sensitivity of HVS to distortions in different spatial frequencies.

Similarly, MSEw overestimates noise level for non eccentricity noise (#14) and for comfort noise (#20) due to inability to account for masking effects.

5. CONCLUSIONS

The paper presents possibilities of verifying quality metrics using TID2013 using MSE and its “derivatives” as examples. The images in TID2013 that have invisible distortions have been marked and this has allowed to evaluate detectability of such images using the considered metrics. Modified versions of SROCC and KROCC are proposed that take into consideration errors of MOS. A way to determine a contribution of a given type of distortion to inadequateness of a given metric to human perception of image quality is proposed.

6. REFERENCES

[1] B.W. Keelan, *Handbook of Image Quality*. Marcel Dekker, Inc.: New York, 2002.

[2] 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, pp. 30-45, 2009.

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

[4] M.G. Kendall, *The advanced theory of statistics. Vol. 1*, London, UK, Charles Griffin & Company limited, 1945.

[5] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. Jay

Kuo, “A New Color Image Database TID2013: Innovations and Results”, *Proceedings of ACIVS, Poznan, Poland*, pp. 402-413. 2013.

[6] 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 5, pp. 2378-2386, 2011,

[7] N. Ponomarenko, O. Eremeev, V. Lukin, K. Egiazarian, and M. Carli, “Modified image visual quality metrics for contrast change and mean shift accounting”, *Proceedings of CADSM*, Ukraine, pp.305-311, 2011.

[8] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, pp. 1398–1402, 2003.

[9] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity”, IEEE Transactions on Image Processing, Vol. 13, Issue 4, pp. 600-612, 2004.

[10] C.X. Ling, J. Huang and H. Zhang, “AUC: a Statistically Consistent and more Discriminating Measure than Accuracy”, Proceedings of 18th International Conference on Artificial Intelligence, pp. 519-526, 2003.