TID2008 – A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics

Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian,
Jaakko Astola, Marco Carli, and Federica Battisti

Abstract— **In this paper, a new image database, TID2008, for evaluation of full-reference visual quality assessment metrics is described. It contains 1700 test images (25 reference images, 17 types of distortions for each reference image, 4 different levels of each type of distortion). Mean Opinion Scores (MOS) for this database have been obtained as a result of more than 800 experiments. During these tests, observers from three countries (Finland, Italy, and Ukraine) have carried out about 256000 individual human quality judgments. The obtained MOS can be used for effective testing of different visual quality metrics as well as for the design of new metrics. Using the designed image database, we have tested several known quality metrics. The designed test image database is freely available for downloading and utilization in scientific investigations.**

Index Terms—**Visual quality metrics, HVS, image denoising, test image databases**

I. INTRODUCTION

The quality evaluation of digital images is critical in many applications of image processing [1, 2]. In the current connected world, many users share and deliver multimedia data. The overall communication process includes manipulation, processing, storing, and transmission over (noisy) channels. Although there have been great improvements in compression and transmission techniques, each stage of processing may introduce perceivable distortions [3, 4]. For example, blocking, ringing, and blurriness are only few of the artifacts that a lossy compression algorithm introduces in an image.

The visibility and annoyance of these impairments are related to the human perception of the quality of the received/processed data. To reduce these impairments it is important to quantify the quality degradations occurred during the processing chain, to maintain, to control, and possibly to enhance the quality of the digital data. To these aims it is crucial to have an effective image quality metric.

Availability of a quality metric adequate to Human Visual System (HVS) is strongly desirable for such applications like image and video lossy compression, quality control of printing and scanning devices, etc.

In general, the quality evaluation should consider the particular application context [5]. For example, when the end user of the image-based system is a human being, the metric used for assessing the overall system effectiveness should take into account the impact of HVS.

Many image quality metrics try to match the HVS [6-9]. All these metrics are in some sense heuristic. However, currently there are no reliable mathematical models for the HVS resulting in the impossibility of defining an optimum metric perfectly matching the HVS. Therefore, a challenging task is the evaluation of the correspondence of visual quality metrics with HVS using some methods of quantitative analysis. Usually this is performed using databases of test images for which the mean opinion scores (MOS) of image quality have been experimentally collected [10]. The methodology for database creation and experimental tests carrying out directly influences accuracy and reliability of quantitative analysis.

This paper presents a new image database TID2008 which is currently up to authors' knowledge the world largest according to the number of test images and types of distortions taken into account [11, 12]. We perform comparison of TID2008 and its closest analog, LIVE Database [10]. The results of verification of many image visual quality metrics using TID2008 are analyzed.

The paper is organized as follows. Peculiarities of using visual quality metrics in digital image processing and requirements to test image databases that stem from these peculiarities are considered in Section II. Section III is devoted to the description of the image database we have created (the set of images, types of distortions, details of generating images with selected types and levels of distortions). In Section IV we present the performed experiments. Analysis of the obtained results is given in Section V. Section VI deals with analysis of known metrics efficiency with exploiting TID2008. Finally, Section VII describes how to get the database at user's disposal, how to use it for user's own purposes, etc.

II. Peculiarities of Using Quality Metrics in Digital Image Processing

According to peculiarities of using visual quality metrics, all methods of digital image processing can be divided into two classes. The first class involves methods of image and video lossy compression for which such metrics can be exploited to control compression quality (Fig. 1).

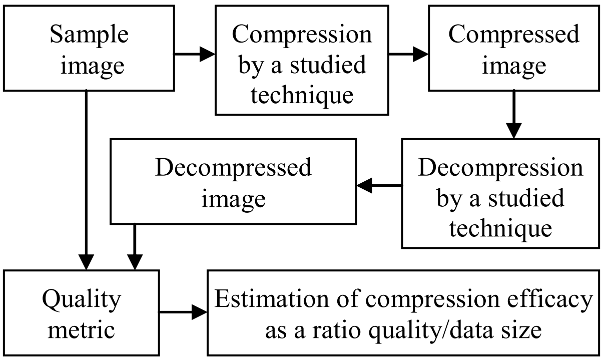

Sample image
Compression by a studied technique
Compressed image
Decompressed image
Decompression by a studied technique
Quality metric
Estimation of compression efficacy as a ratio quality/data size

Fig. 1. Block diagram of the process of efficiency verification for a given lossy compression technique

In this case, a quality metric can be used both in the design of the compression block and in the overall performances evaluation. In the latter case, a metric value calculated for a decoded image can be used in the tuning phase of the parameters in the coarse-to-fine compression schemes. Then if, for example, an obtained value of quality metric is inappropriate, an image can be compressed with better quality with smaller quantization step or, equivalently, with larger bit rate.

Although many existing models of HVS are able to consider such parameters as the distance of human eyes to a monitor, the monitor size, etc., it is often useful to have a quality metric independent from the knowledge of these parameters to mimic the common restitution environment. Since monitor characteristics and observer-monitor distance are highly varying parameters, in our test design we prefer to consider different situations by averaging several viewing conditions.

Another class of digital image processing techniques involves a variety of methods such as image filtering, reconstruction, inpainting, etc. For this class, image visual quality metrics are used only in the process of a method design and evaluation of its efficiency.

It is well known that practically it is impossible to define the optimum filtering technique due to the non-stationary nature of processed 2D data. Therefore, statistical verification of efficiency has to be employed. Fig. 2 presents a block diagram of quantitative verification of efficiency evaluation for image filtering methods.

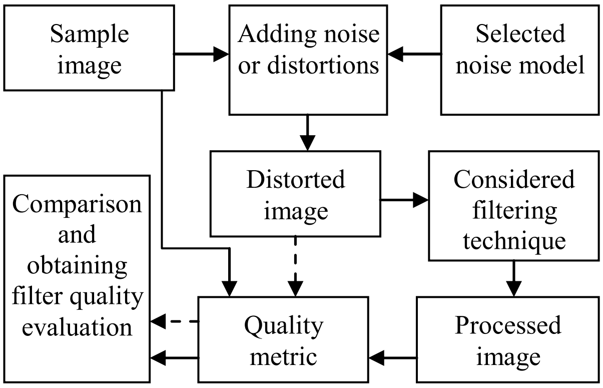

Sample image
Adding noise or distortions
Selected noise model
Distorted image
Considered filtering technique
Comparison and obtaining filter quality evaluation
Quality metric
Processed image

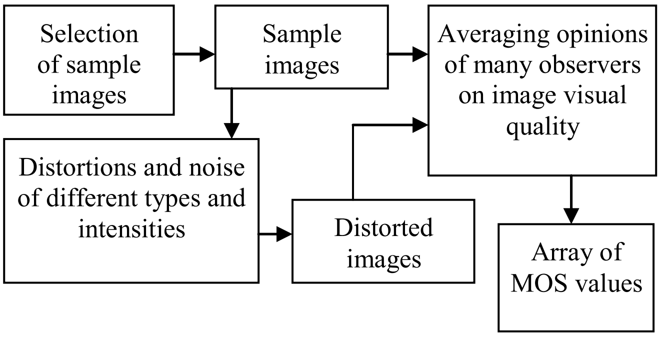
Fig. 2. Block diagram of the process of efficiency verification for a given method of image filtering

As a first step, a test image (or a set of test images) presenting *good* quality is selected. Then, according to the chosen model of noise or distortion, a noisy version of the image (images) is obtained and processed by a designed filter. The obtained output image is ”compared” to the corresponding original image using considered quality metric. A value of the same metric is calculated for the noisy (distorted) image as well. By comparing the scores of these metrics it is possible to address the effectiveness of the designed filtering technique.

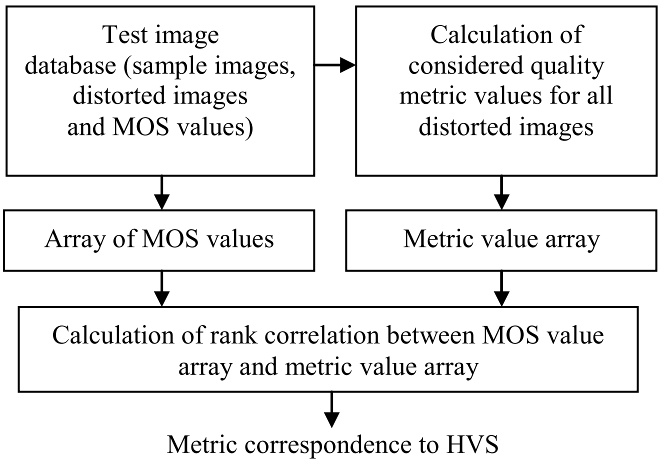
Requirements for quality metrics applied to the second class of digital image processing techniques are equivalent to those for image compression case. In this case as well we intend to evaluate the metrics for some averaged conditions of image visualization. Note that metrics used in block diagrams in Figures 1 and 2 are full reference ones, i.e., they are calculated for pairs of a reference and the corresponding distorted images.

As can noticed from the few examples previously discussed, the availability of a good visual quality metric is needed to adequately assess efficiency of an image processing method or visual quality of compressed images. To understand the relation between the given metric and the HVS, the most reliable way is to exploit some specially created test image database.

Fig. 3 shows the use of an image database in testing visual quality metrics.


Selection of sample images
Sample images
Averaging opinions of many observers on image visual quality
Distortions and noise of different types and intensities
Distorted images
Array of MOS values

a)


Test image database (sample images, distorted images and MOS values)
Calculation of considered quality metric values for all distorted images
Array of MOS values
Metric value array
Calculation of rank correlation between MOS value array and metric value array
Metric correspondence to HVS

b)

Fig. 3. Creation (a) and use (b) of test image database for verification of visual quality metrics

The designed image database (Fig. 3, a) consists of a set of sample (reference) images, a corresponding set of distorted images and an array of MOS values obtained for each distorted image. When a metric is tested, its values are computed for each distorted image in a database and the obtained results are ”compared” to the corresponding MOS values. To avoid scaling problems the rank correlations are often used, e.g., Spearman and Kendall ones [13]. Larger values of rank correlations (the maximal value equals to unity) indicate better correspondence of a given metric to image quality assessment by humans.

Note that after preliminary fitting metric values and MOS, it is, in general, possible to apply the conventional Pearson correlation instead of rank correlation. Meanwhile, it is important to underline that the quality of fitting may reduce the accuracy of assessing a metric correspondence to HVS.

The image databases to be used for the chosen application should satisfy several requirements. The main constraint is that the test images should reflect the HVS peculiarities and contain non-trivial images for visual quality evaluation in order to effectively retrieve the advantages and the drawbacks of all tested quality metrics. Thus, it is possible to define the following requirements for the test image database:

- it should include images with considerably different characteristics: percentage of homogeneous regions, details and textures, various texture characteristics, etc.;
- for each HVS feature, the database has to contain, at least, one distortion type that allows to estimate how this feature influences image visual quality;
- it is desirable that the database will contain image distortions typical for practice that originate due to compression, denoising, data transmission errors, etc;
- the images in the database should be challenging for visual quality estimation, however 1) the number of distortion levels should not be large, 2) the number of situations when all metrics evidence in favor of a given image should not be large.

Furthermore, it has been stated above that tested metrics are to be oriented on some average conditions (parameters) of image visualization. Then, experiments intended on obtaining observers’ opinions (MOS values) in creating and exploiting test image databases have to be carried out with reasonable variation of image visualization conditions.

Note that under this assumption an image database can not be used for a design and verification of HVS models since creation (design) of HVS models requires strict control of visualization and observation conditions (parameters). However, if images are visualized and analyzed in slightly varying conditions, this, to our opinion, provides the best verification of quality metrics if they are intended for visual quality assessment in a priori unknown and variable conditions of visualization and observation.

III. Description Of The Proposed Image Database

The quality of any image database strictly depends on the reference images that are used. The main strategy is to select the images that represent a wide variety of scenes. That is, the images in the database should present different textural characteristics, various percentage of homogeneous regions, edges, and details. In this sense, the Kodak test set [14] can be considered as a good trade off between the abovementioned requirements. This set is the basis of the LIVE database [10] and we have also used it. Recall that the Kodak test set contains 24 images (see Fig. 4).

Besides, we have synthesized and added to TID2008 one artificial image (Fig. 4) that has different texture fragments and objects with various characteristics. The motivation of including this image into TID2008 was to provide adequate testing for metrics intended to work with such kind of images.

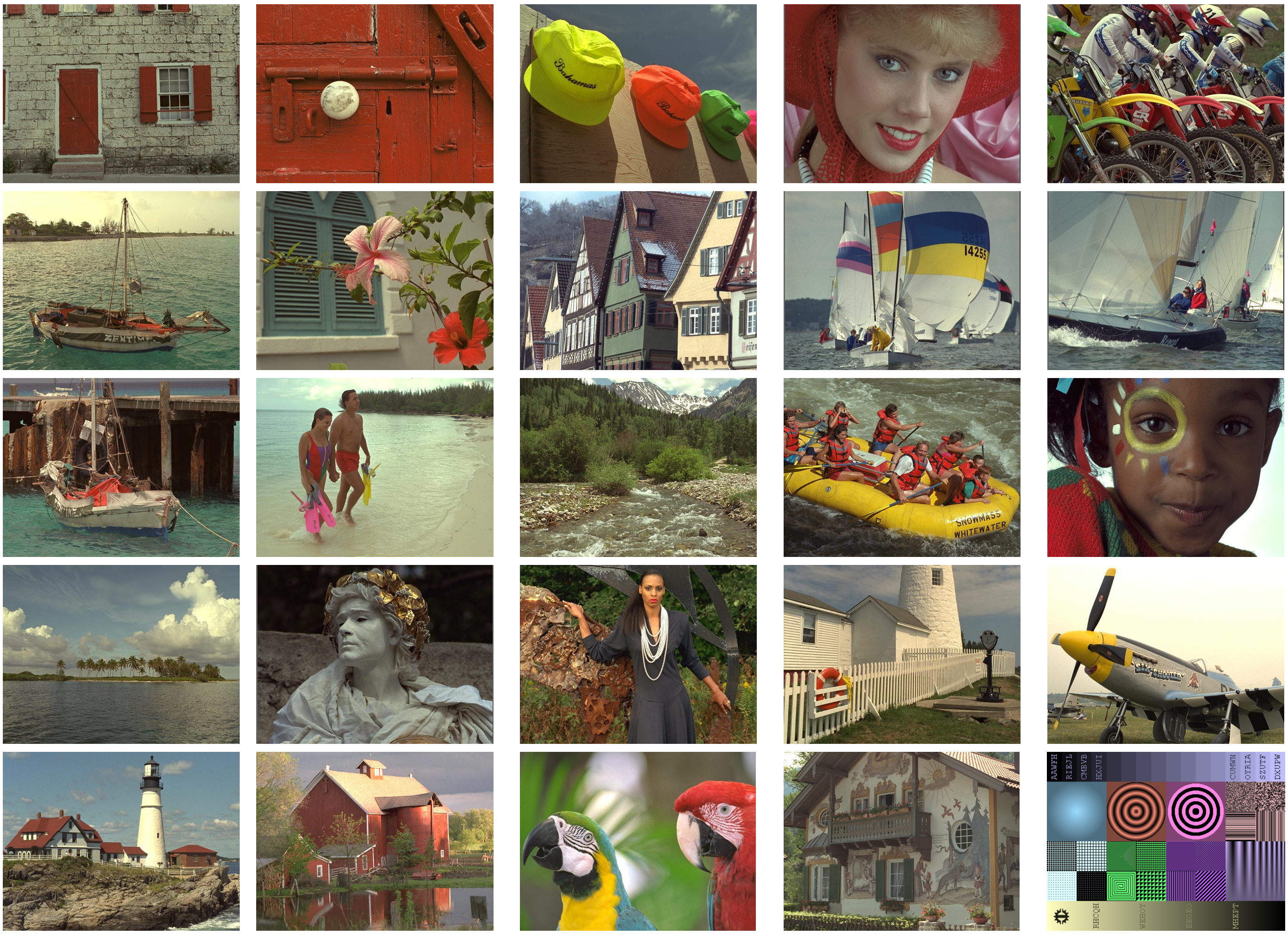


Fig. 4. Reference images (1-24) of our database formed using the Kodak test set and the 25-th reference image synthesized by us (each 512x384 pixels, 24 bit per pixel)

Table I. Types of distortions used in our image database

№	Type of distortion (four levels for each distortion)	Correspondence to practical situation	Accounted HVS peculiarities
1	Additive Gaussian noise	Image acquisition	Adaptivity, robustness
2	Additive noise in color components is more intensive than additive noise in the luminance component	Image acquisition	Color sensitivity
3	Spatially correlated noise	Digital photography	Spatial frequency sensitivity
4	Masked noise	Image compression, watermarking	Local contrast sensitivity
5	High frequency noise	Image compression, watermarking	Spatial frequency sensitivity
6	Impulse noise	Image acquisition	Robustness
7	Quantization noise	Image registration, gamma correction	Color, local contrast, spatial frequency
8	Gaussian blur	Image registration	Spatial frequency sensitivity
9	Image denoising	Image denoising	Spatial frequency, local contrast
10	JPEG compression	JPEG compression	Color, spatial frequency sensitivity
11	JPEG2000 compression	JPEG2000 compression	Spatial frequency sensitivity
12	JPEG transmission errors	Data transmission	Eccentricity
13	JPEG2000 transmission errors	Data transmission	Eccentricity
14	Non eccentricity pattern noise	Image compression, watermarking	Eccentricity
15	Local block-wise distortions of different intensity	Inpainting, image acquisition	Evenness of distortions
16	Mean shift (intensity shift)	Image acquisition	Light level sensitivity
17	Contrast change	Image acquisition, gamma correction	Light level, local contrast sensitivity

All images in our database are of size 512x384 pixels. This choice has been suggested by unification purpose (images in the Kodak test set have non-equal sizes) and for a convenience of carrying out subjective experiments (see Section IV). All images of fixed size have been obtained by cropping selected fragments from the original images of the Kodak test set without any scaling and/or rotation operations.

Table I presents the distortions modeled in our image database. Additive zero-mean noise is often present in images [2] and it is commonly modeled as a white Gaussian noise. This type of distortion is included in most of studies of quality metric effectiveness. This type of distortion is, probably, one of few cases when metrics MSE and PSNR present a good match with the HVS.

The distortion type 2 (noise is non-uniformly distributed between color components modeled in the color space YCbCR [15]) has been added to test the quality metric correspondence to known property of HVS to not equally perceive distortions in brightness (luminance) and color (chrominance) components.

Quite often additive noise cannot be considered as spatially uncorrelated (white noise). Thus, consideration of spatially correlated noise allows, first, to check metrics' correspondence to inherent spatial frequency sensitivity of HVS. Second, such kind of noise is present in the important class of color images created by modern digital cameras [16]. Recall that an image taken from a digital camera matrix is mosaic [17]. Such images are corrupted by noise with rather complicated statistical properties [18] but originally noise is practically spatially independent. However, in the process of image converting from original mosaic to further used raster form data are subject to nonlinear interpolation and noise becomes spatially correlated. Noise removal and image compression for this application are quite complex tasks [16], [19], [20]. Currently a great interest is observed in the design of effective methods to solve aforementioned tasks and it is important to adequately evaluate image visual quality before and after processing.

Low-pass spatially correlated noise is not the only case of additive noise that is not white. Masked noise and high frequency noises are other types of distortions that allow analyzing metrics' adequateness with respect to local contrast sensitivity and spatial frequency sensitivity of HVS. Such types of distortions are typical for a wide class of practical tasks like lossy image compression and, especially, digital watermarking [21], [22]. As it was demonstrated by some studies [7], many known quality metrics, unfortunately, do not take these peculiarities of HVS into account well enough.

Impulse noise (we have used a typically used model of uniformly distributed impulse noise [23]) arises, in particular, due to coding/decoding errors in data transmission.

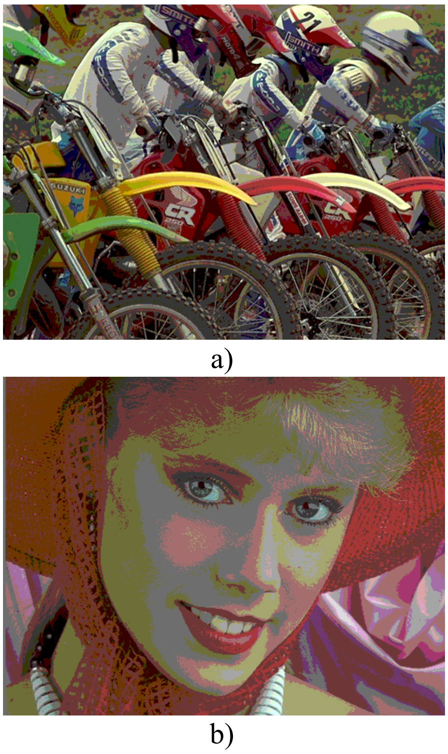
The task of its removal is of great interest during the last three decades [24]. Thus, to our opinion, the presence of images affected by impulse noise in the database is necessary. This might assist to adequately evaluate effectiveness of methods for impulse noise removal, image inpainting [25], etc. Besides, the use of such images might help in assessing how the tested metrics account for such property of HVS as robustness to impulses. It has been proven that for small probabilities of impulse noise humans are able to quite easily intuitively recover the values for pixels corrupted by spikes using neighbor pixels.

Quantization noise has not received too much attention in image visual quality evaluation, although this distortion is quite often met in practice and it allows to estimate quality of the metrics' adequacy with respect to several peculiarities of HVS. Quantization noise characterized by the same PSNR can be almost not be noticed in highly textured images (Fig. 5 (a)) while being very noticeable in images with few textured areas (Fig. 5 (b)). Most of known metrics do not adequately assess visual quality of images subject to the considered type of distortion.

Gaussian blur is also considered in the proposed database since it is an important type of distortions often met in practical applications and frequently included in studies dealing with visual quality metrics [10].

Other important type of distortions studied recently [24] are residual distortions resulted after applying different denoising procedures (filters). Image filtering constitutes a class of practical tasks for which it is necessary to have an appropriate tools to evaluate visual quality of filtered image.

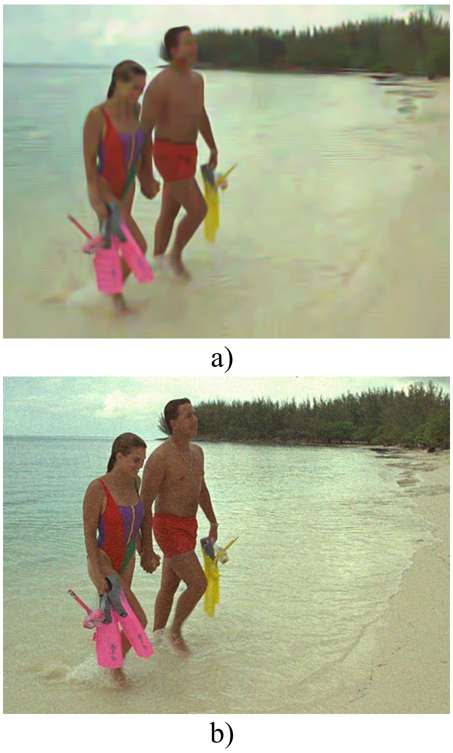
Unfortunately, it often happens that a filtered image presents a higher PSNR value (2-3 dB higher) than the original one, but, at the same time, a processed image looks perceptually worse than the corresponding noisy original. Thus, we have included into our database images for which original additive i.i.d. Gaussian noise is suppressed by one of the best state-of-the-art filter [26] based on 3D Discrete Cosine Transform (DCT).



a)

b)

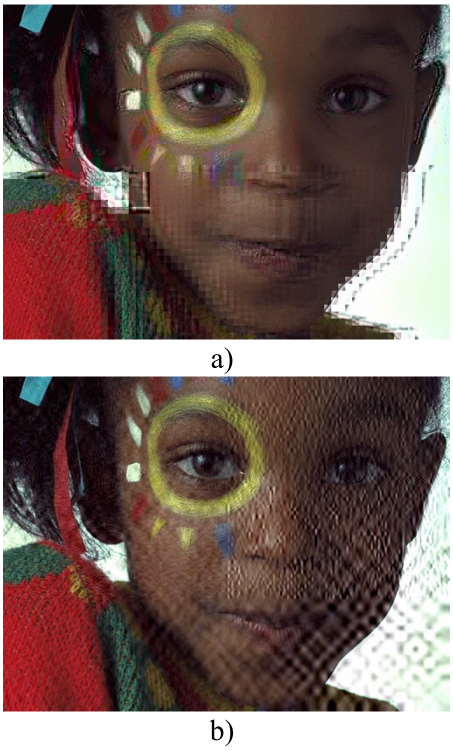
Fig. 5. Two test images distorted by quantization noise: a) the highly textured image, PSNR=24.35 dB, b) the image containing much less texture, PSNR=24.21 dB



a)

b)

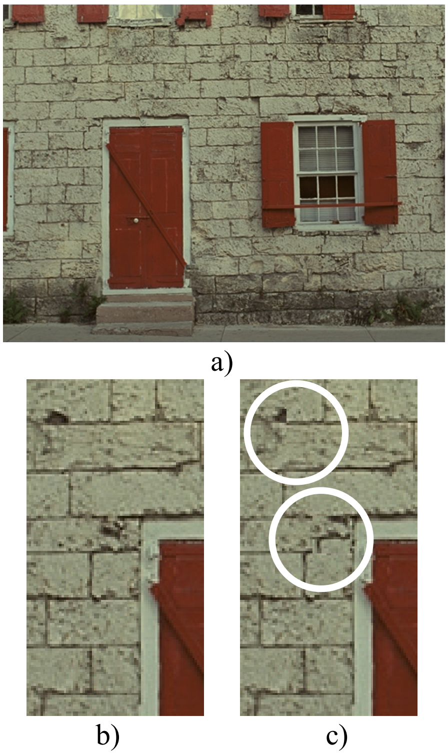
Fig. 6. Comparison of visual quality a) after filtering out additive noise, PSNR=28.19 dB, b) original noisy image corrupted by additive i.i.d. Gaussian noise, PSNR=26.99 dB



a)

b)

Fig. 7. Image decoded with errors due to unreliable data transmission line: a) for the standard JPEG, PSNR=24.05 dB, b) for the standard JPEG2000, PSNR=23.98 dB

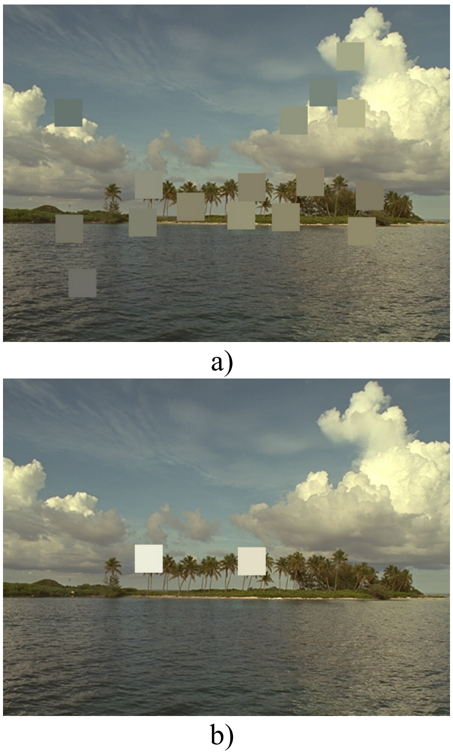


a)

b)

c)

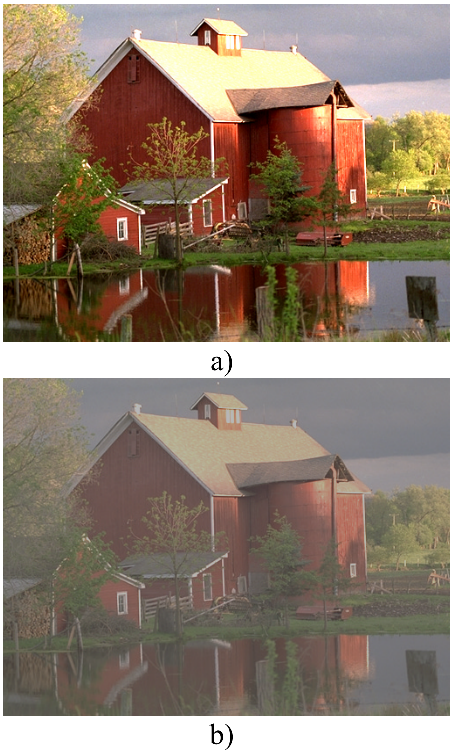
Fig. 8. An example of non-eccentricity distortions: a) distorted image, PSNR=27.0 dB, b) enlarged fragments of the reference image, c) the same fragment with introduced distortions (the corresponding places



a)

b)

Fig. 9. An example of block-wise distortions of different intensity: a) 16 blocks, PSNR=26.49 dB, b) 2 blocks, PSNR=25.90 dB



a)

b)

Fig. 10. An example of contrast change: a) to larger contrast, b) to smaller contrast

are marked by white circles)

According to [26], this filtering approach has produced the best visual quality of filtered images among several considered effective filters.

Fig. 6 gives an example of the processed image that can be compared to original image corrupted by Gaussian additive noise. As it can be seen, although the processed image is characterized by a larger PSNR, residual noise after filtering and distortions that are inevitably introduced by any filter lead to sufficient visual artifacts.

Similarly to LIVE database, images distorted with lossy compression (JPEG and JPEG2000) have been included into our database. The tasks of evaluating distortions for lossy image compression techniques are of great interest. Besides, we have included into our database the images compressed by JPEG or JPEG2000 and decoded with errors in data transmission channels. Decoding errors have been modeled in such a way that required PSNR has been provided for decoded images. Quite often distortions induced by such errors are almost invisible thanks to their non-eccentricity. Fig 7 presents two examples of distortions due to transmission/decoding errors. Distorted fragments might occur to be similar to original texture and/or color of surrounding fragments and due to peculiarities of HVS a human might not notice such distortions. To our opinion, the use of images for which the considered distortions are modeled allows to predict the ability of the tested quality metrics to include this HVS feature into account.

HVS usually is not sensitive to non-eccentricity distortions: by considering this behavior in applications like lossy image and video compression might lead to considerable quality improvement of compressed images with the same compression ratio. Therefore, we have decided to include into the database the images distorted by a specific type of artifacts modeled by us and called “non eccentricity pattern noise”. Such distortions have been modeled in the following way. A small image fragment of size 15x15 pixels has been randomly taken in a reference image and it has been copied instead of other fragment located nearby (at distance of few pixels). This operation has been repeated several times until a required PSNR is approximately provided. Note that without having the corresponding reference image, by analyzing the image in Fig. 8,a it is even difficult to localize the distortion.

For instance, if an image is corrupted by impulse noise, a human being can easily detect the corresponding pixels. But in the considered case it is frequently difficult to identify compact distortions of rather large size of 15x15 pixels. This demonstrates the property of HVS to discard non-eccentricity distortions.

Another specific type of distortions modeled by us and added to TID2008 are the so called local block-wise distortions of different intensity. We suppose that in case of compact impulse-like distortions, HVS does not react to distortion on single pixel but mainly to an area (percentage of pixels) that is a subject to distortions. Distortions have been modeled in such a way that blocks of size 32x32 pixels having arbitrary random color are randomly placed in important areas of an image.

For the first level of distortions, 16 blocks having color slightly different from the mean color of replaced fragment have been added (see Fig. 9,a). For the second level of distortions, the amount of such blocks was 8 but their color differs more from mean color of replaced fragment. For the third and fourth levels, four and two blocks have been replaced, respectively. However, for these blocks their color differs even more essentially from the mean colors of the corresponding replaced fragments (Fig. 9,b). Color and intensity differences have been adjusted in such a manner that irrespective to the number of blocks approximately the same PSNR has been provided. Example in Fig. 8 shows that the image corrupted by two blocks is perceived as having better visual quality (although it has smaller PSNR) than the image distorted by 16 blocks. This has been confirmed by experiments (see Section IV). Most probably, such assessment (decision) is explained by HVS inability to “retrieve” lost information in places distorted by blocks irrespective to their color. Then it seems that from image perception point of view it is better if total area of such block-wise distortions is smaller whilst a degree of such distortions is of less importance.

Finally, we have added into our database images for which mean shift and contrast change distortions have been modeled. Importance of these distortions has been demonstrated in [10]. Mean value shifting and contrast changing have been done with respect to images as a whole. For each distortion of mean and contrast two changes to smaller and two changes to larger values have been simulated (see images in Fig. 10).

As it was mentioned above, we have set four levels for all types of distortions. For almost all types of distortions, the corresponding levels of PSNR are about 30 dB, 27 dB, 24 dB, and 21 dB (very good quality, good quality, poor quality, and bad quality). On one hand, such number of distortion levels for 25 reference images allows to take into account all range of subjective quality of distorted images from “excellent” to “very bad”. On the other hand, four levels do not create too many simple combinations of image pairs at their quality comparison stage (see Section II).

It is possible to assert that in the presented database, to simulate some types of distortions we have exploited either some very particular or some quite simple models of distortions. For example, this relates to impulse noise for which different models exist [23] or for blur that can be characterized by point spread functions with a wide range of parameters [19]. In this sense, our intention was, on one hand, to model more types of possible distortions than in other databases of distorted images. On the other hand, we used quite typical and simple models that allowed obtaining distorted images in an easy way..

IV. Experiments Description

In the performed experiments, a large group of observers (volunteers) have been evaluating visual quality of distorted images in the database. As a result, MOS values have been obtained.

There are different methodologies that can be used to evaluate the quality of an image [7, 10, 27]. Depending on a chosen strategy, the observers have been asked to evaluate the absolute quality of an image or its similarity to a reference. In both cases the subject evaluation is expressed with a grading scale that can be continuous or discrete, categorical or numerical (see Fig. 11).

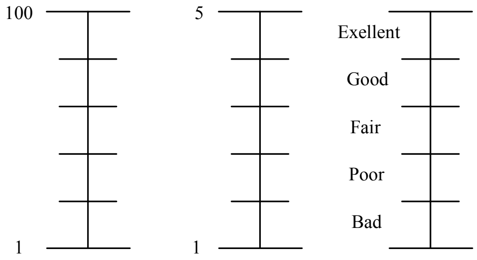
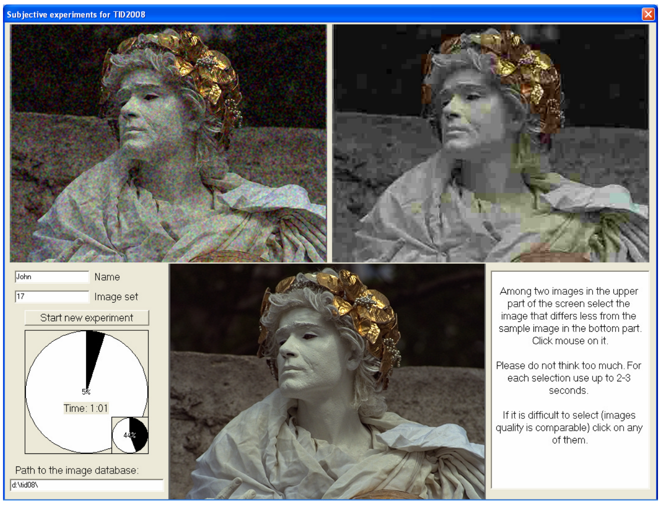


Fig. 11. Examples of possible grading scales

For example, in [10] five gradations have been used and they have been described through the five categories: “Bad”, “Poor”, “Fair”, “Good” and “Excellent” (or, equivalently, with a score from 1 to 5). However, it is often difficult for an observer to assign scores to the distorted images. This has been stated by many persons participated in experiments carried out according to this methodology. In fact in many cases, an observer has evaluated quality of a distorted image “A” as “Bad”, but later he/she has to evaluate quality of another distorted image “B” that is even worse. But there is no gradation worse than “Bad” in the used scale. This leads to insecurity of the observer and in the willingness to change the previously given grade. The grade change is often not permitted by common assessment systems. To simplify the evaluation procedure, the observers undergo a training phase in which they can see some examples of the distortions that are present in the test set; this helps in getting an idea of what is “Bad” or “Excellent” quality [28].

When designing MOS for TID2008, we have used another methodology for carrying out the subjective tests. At monitor, the reference image (in the lower part) and a pair of distorted images (in the upper part) are simultaneously presented (see an example in Fig. 12).


Subjective experiments for TID2008

John Name

17 Image set

Path to the image database: d:\tid08\

Among two images in the upper part of the screen select the image that differs less from the sample image in the bottom part. Click mouse on it.

Please do not think too much. For each selection use up to 2-3 seconds.

If it is difficult to select (images quality is comparable) click on any of them.

Fig 12. Screen-shot of the software used in experiments

Each observer was asked to select a distorted image (between two ones) that differs less from the reference one. After the first selection, two different (new) distorted images appear in the upper part of screen.

Such approach (methodology of comparisons) has been proven to be less annoying for experiment participants although, according to opinions of some researchers [29], it produces less accurate estimates of MOS. For TID2008, we have derived estimates of MOS accuracy for the considered approach and compared them to accuracy of MOS estimation produced by conventional MOS derivation for estimation of observers’ quality. These estimates are presented in Figures 13 and 14.

Let us describe the process how experiments have been performed to obtain the MOS. Each observer in one experiment has carried out distorted image quality assessment for only one reference image (68 distorted images). Average time for one experiment takes about 13.5 minutes. Such approach (short-time experiments) was used in order to reduce the load for each participant to the experiment. According to recommendations given in [27], the time of accomplishing one experiment by each observer should not exceed 30 minutes.

However, the approach followed in [10], when each experiment was carried out separately for each type of distortion (with the same purpose to decrease the time of experiment) may lead to over-learning to a given type of distortion.

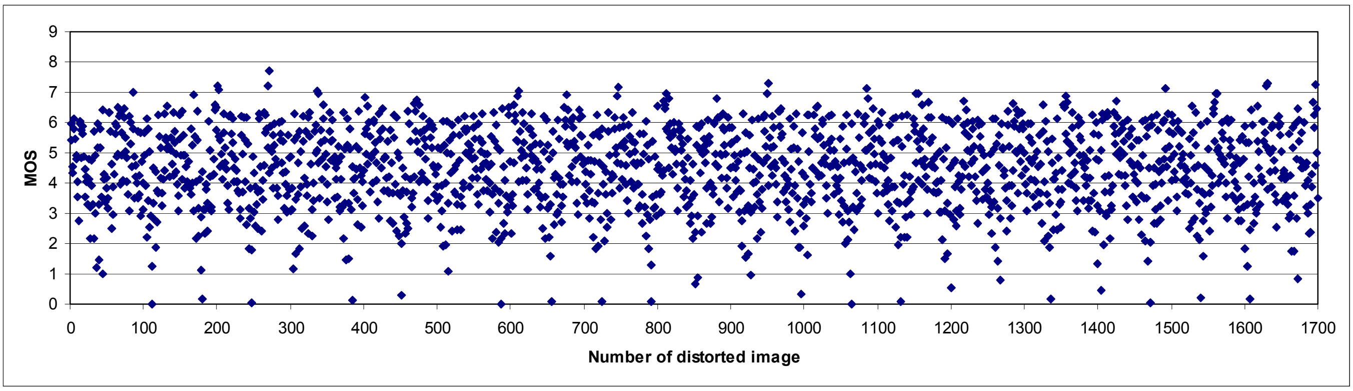


Fig. 13. MOS for TID2008 images

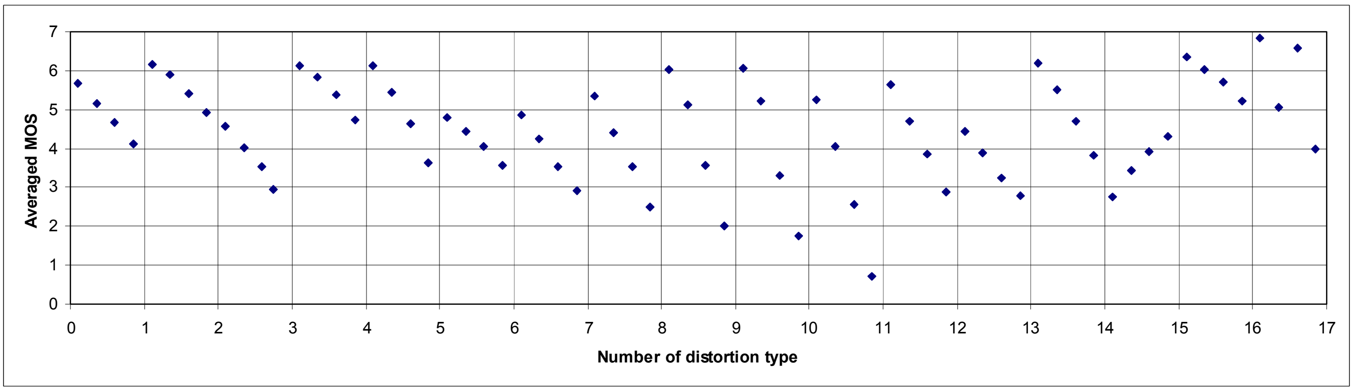


Fig. 14. MOS averaged for all reference images

Totally, more than 800 observers of different cultural level (students, tutors, researchers) from three countries (Finland, Italy, Ukraine) have participated to the experiments. More than 200 observers have carried out experiments remotely, via Internet. Other observers have performed experiments in classes, under control and supervision of tutors. Each observer (including those involved remotely) was preliminarily instructed and trained on a set of distorted images before carrying out the actual experiments.

According to requirements presented in Section II, visualization and observation conditions varied in reasonable limits to be comfortable for each participant. Different monitors were used, both LCD and CRT, mainly 19” with the preset resolution 1152x864 pixels.

Each observer in one experiment has been asked to select an image with better quality between two for 306 times. Because it is practically impossible to carry out pair-wise comparisons of each of the 68 distorted images to another distorted image, we exploited Swiss competition principle (used in chess tournaments). At the first step, 68 images were divided into 34 random pairs and were shown to an observer. An image pointed as a «winner» got one point. At the next steps, only representatives of the same point groups were used for comparison, making totally, 9 tours (34 x 9 = 306). Therefore, each image could get from 0 to 9 points as a result of experimental tests. These amounts of points represent quantitative evaluation of image quality obtained from a given observer.

Evaluations of image quality provided by different observers might have abnormal evaluations (e.g., due to wrong clicks, or to low attention level in some observer). The validity of the subjective test results was verified by a screening of the results performed according to Annex 2 of ITU-R Rec. BT.500 [27]. We used the same methodology as in [10]. About 2% of abnormal evaluations (outliers) have been removed and 14 abnormal experiments out of 838 (about 1.7%) have been neglected.

For each reference image, about 33 evaluations of distorted image visual quality have been done using a scale from 0 to 9 with the mean close to the true value of MOS and variance about σ^2=0.63 (relative variance is about 0.031). The final values of MOS for each image have been obtained by averaging all quality evaluations for a given image. Thus, most of MOS values obtained by exploiting TID2008 are characterized by Gaussian errors and they correspond to an ideal quality metric that corresponds to HVS with σ^2 approximately equal to 0.02 (0.63/33). For the considered scale 0..9 this is quite small variance and this evidences in favor of high accuracy of the obtained MOS estimates.

In order to compare our results with the conventional quantitative evaluation of image visual quality [10] we have carried out a separate experiment. 60 observers have evaluated the quality of distorted images giving a grade in a scale from 1 to 5. The experiment has been done only for the first reference image. 2% of outliers have been removed and 10% of abnormal experiments have been rejected. Variance of the obtained MOS estimates for the scale from 1 to 5 was 0.4 (for the scale 1..100 this is equivalent to $\sigma^2 = 250$). Relative variance (the ratio of variance to squared mean) for this approach is 0.083. This is almost three times larger than for the approach used by us in earlier experiments by exploiting TID2008. Thus, the methodology we have used to perform the experimental tests shows better accuracy. Also note that Spearman correlation for MOS obtained for the first original image family (68 distorted images) for both methodologies of the experiment is equal to 0.97. This correlation factor is very high and it demonstrates that both methodologies lead to similar results. This shows that both methodologies can be used and they differ only by convenience for observers and by the provided accuracy of MOS estimation.

One more indirect evidence in favor of the used methodology in the way the experiments have been performed in TID2008 is the high correlation value between the data obtained for observers from different countries (see Table II).

TABLE II. CORRELATION BETWEEN MOS FOR GROUPS OF OBSERVERS IN THREE PARTICIPATING COUNTRIES

Countries	Spearman correlation
Finland (251 observers) - Italy (150 observers)	0.93
Finland (251 observers) - Ukraine (437 observers)	0.96
Italy (150 observers) - Ukraine (437 observers)	0.93

Note that more than 200 observers in Finland have carried out experiments via Internet. Being a distance-based methodology, this could have lead to larger number of outliers. However, high correlation for the results obtained in Finland with the results in Italy and Ukraine (where experiments have been performed in class-room) show confidence in the overall scores.

Fig. 13 presents MOS for all 1700 test images in TID2008 and Fig. 14 shows averaged MOS for each type and level of distortions.

V. COMPARISON OF TID2008 AND LIVE DATABASES

At the moment, TID2008 is the largest database of distorted images intended for verification of full-reference quality metrics. Table III presents the main parameters and characteristics allowing the comparison of TID2008 to its nearest analog LIVE Database [10].

TABLE III. COMPARISON CHARACTERISTICS OF LIVE DATABASE AND TID2008 DATABASE

N	Main characteristics	Test image database	
		LIVE Database	TID2008
1	Number of distorted images	779	1700
2	Number of different types of distortions	5	17
3	Number of experiments carried out	161 (all USA)	Totally 838 (437 - Ukraine, 251 - Finland, 150 - Italy)
4	Methodology of visual quality evaluation	Evaluation using five level scale (Excellent, Good, Fair, Poor, Bad)	Pair-wise sorting (choosing the best that visually differs less from original between two considered)
5	Number of elementary evaluations of image visual quality in experiments	25000	256428
6	Scale of obtained estimates of MOS	0..100 (stretched from the scale 1..5)	0..9
7	Variance of estimates of MOS	250*	0.63
8	Normalized variance of estimates of MOS	0.083*	0.031

Estimates marked by "*" in Table III, are obtained as the result of experiments described in the previous Section.

The main advantage of TID2008 with respect to LIVE Database is that TID2008 accounts for 17 different types of distortions and, thus, covers more practical applications and known peculiarities of human visual system (HVS). LIVE Database deals with only five types of distortions for which most known metrics tested using LIVE Database commonly have quite good correspondence to HVS. In turn, TID2008 allows carrying out more detailed analysis of quality metrics indicating their drawbacks and demonstrating prospective ways of further investigations and design.

Besides, the MOS values for TID2008 are more accurate than for LIVE Database. For comparable number of visual quality evaluations for each image, relative variance for TID2008 is almost three times smaller than for LIVE Database.

VI. COMPARATIVE ANALYSIS OF QUALITY METRICS

Quality metric verification using TID2008 can be done using all MOS as well as for particular subsets of TID2008. A subset may include one or several types of distortions. Table IV shows subsets used below for verification of quality metrics (distortions that belong to a given subset are marked by +).

TABLE IV. DISTORTION TYPES AND CONSIDERED SUBSETS OF TID2008

№	Type of distortion	Noise	JPEG	Exotic	Actual	Full
1	Additive Gaussian noise	+	-	-	+	+
2	Different additive noise 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	-	-	-	-	+

We have evaluated correspondence of HVS to the following 18 metrics (quality indices): MSSIM [8, 30], VIF [31, 30], a pixel based version of VIF (VIFP) [31, 30], VSNR [9, 30], PSNR-HVS (PSNRHVS) [28], PSNR-HVS-M (PSNRHVSM) [7], SSIM [6], NQM [32, 30], UQI [33], XYZ [34], LINLAB [35], IFC [36, 30], WSNR [37, 30], DCTUNE [38], SNR [30], MSE [30], PSNR [30] and PSNR calculated for only brightness (intensity) component of color images (PSNRY). Table V presents the values of Spearman correlation for the considered 18 metrics and the subsets used in TID2008. Similarly, Table VI contains the corresponding values of Kendall correlation factors. The first row of both Tables presents correlations between obtained MOS and "ideal" MOS that could be provided if the number of experiments approaches to infinity.

Here we would like to emphasize that most metrics analyzed in this paper are oriented on application for grayscale images. Thus, they have been calculated with respect to intensity images of color images used in TID2008. Meanwhile, TID2008 contains three types of distortions (namely, numbers 2, 10, and 12) that are not uniformly distributed between color (RGB) components. Thus, TID2008 can be used to verify both types of metrics, those that take and those that not take into consideration color information. Three best metrics producing the greatest correlations for each subset are marked in bold in Tables V and VI.

TABLE V. SPEARMAN CORRELATIONS FOR THE CONSIDERED METRICS

№	Metric	Noise	JPEG	Exotic	Actual	Full
-	Ideal metric (HVS)	*0.991*	*0.996*	*0.985*	*0.994*	*0.994*
1	MSSIM	0.813	0.957	***0.673***	0.868	***0.853***
2	SSIM	0.856	***0.964***	0.468	0.882	***0.808***
3	VIF	0.820	0.956	0.045	0.841	***0.750***
4	VSNR	0.857	0.930	0.490	0.869	0.705
5	VIFP	0.734	0.949	0.033	0.821	0.655
6	NQM	0.865	0.932	0.517	0.874	0.624
7	UQI	0.526	0.860	0.156	0.677	0.600
8	PSNRHVS	***0.917***	***0.966***	0.541	***0.920***	0.594
9	XYZ	0.848	0.815	***0.679***	0.829	0.577
10	IFC	0.663	0.898	-0.075	0.732	0.569
11	PSNRHVSM	***0.918***	***0.971***	0.518	***0.929***	0.559
12	PSNRY	0.752	0.866	0.630	0.810	0.553
13	SNR	0.712	0.805	0.561	0.760	0.523
14	MSE	0.704	0.877	***0.671***	0.794	0.525
15	PSNR	0.704	0.877	***0.671***	0.794	0.525
16	WSNR	***0.897***	0.949	0.544	***0.900***	0.488
17	LINLAB	0.839	0.906	0.604	0.847	0.487
18	DCTUNE	0.864	0.933	0.556	0.860	0.476

TABLE VI. KENDALL CORRELATIONS FOR THE CONSIDERED METRICS

№		Noise	JPEG	Exotic	Actual	Full
-	HVS	*0.921*	*0.947*	*0.902*	*0.933*	*0.935*
1	MSSIM	0.609	0.818	0.478	0.675	***0.654***
2	SSIM	0.658	***0.828***	0.311	0.691	***0.605***
3	VIF	0.634	0.814	0.092	0.657	***0.586***
4	VSNR	0.665	0.764	0.372	0.677	0.534
5	VIFP	0.536	0.806	0.082	0.631	0.495
6	PSNRHVS	***0.751***	***0.837***	0.385	***0.750***	0.476
7	NQM	0.673	0.766	0.349	0.678	0.461
8	PSNRHVSM	***0.752***	***0.847***	0.364	***0.765***	0.449
9	UQI	0.363	0.666	0.115	0.489	0.435
10	XYZ	0.654	0.633	***0.480***	0.638	0.434
11	IFC	0.477	0.714	0.004	0.542	0.426
12	PSNRY	0.549	0.670	0.452	0.609	0.402
13	WSNR	***0.714***	0.797	0.379	***0.715***	0.393
14	LINLAB	0.652	0.758	0.422	0.665	0.381
15	SNR	0.512	0.604	0.396	0.558	0.374
16	DCTUNE	0.683	0.791	0.379	0.676	0.372
17	MSE	0.501	0.692	***0.488***	0.593	0.369
18	PSNR	0.501	0.692	***0.488***	0.593	0.369

We would like to draw readers' attention to the fact that Spearman correlation values for the metrics PSNR and MSE are equal. If the conventional Pearson correlation is used, then, without fitting, the correlation factor for these metrics might be not equal to unity although they are strictly connected. Then, for increasing correlation of metrics their fitting is needed [10]. Because of this, it is preferable to employ rank correlation that avoids necessity of fitting in the considered analysis. This is also important because a quality of fitting commonly influences accuracy of obtained results.

The data presented in Tables V and VI for the whole image database TID2008 (the set marked as "Full") show that the widely used metrics PSNR and MSE have very low correlation with human perception (correlation factors are about 0.5). Even the best among considered metric MSSIM has correlation with HVS of the order 0.85 whilst it is desirable to provide a Spearman correlation value around 0.99.

For the subset "JPEG", the best Spearman correlation (SC) between MOS and analyzed metrics is provided by our metrics PSNR-HVS and PSNR-HVS-M [7] (SC is about 0.97), slightly smaller SCs are observed for the metrics MSSIM and VIF (about 0.96). Applicability of these metrics for lossy image compression has been also recently pointed out in [39].

For the subset "Noise", the largest values of both Spearman and Kendall correlations (about 0.92 and 0.75, respectively) have been observed for the metrics PSNR-HVS and PSNR-HVS-M as well. The metric Weighted SNR (WSNR) performs for this set rather well (SC is about 0.90 and Kendall correlation equals to 0.71).

For "Exotic" subset (that, in fact, includes different versions of impulsive distortions) even the best metric MSSIM exhibits low values of SC and KC which are the smallest between the considered subsets MOS (SC=0.679, KC=0.488). Surprisingly, just for this subset the metrics MSE and PSNR perform better than other metrics (according to Kendall correlation). This indirectly shows that till now practically no attention in metric design and analysis has been paid to consider distortions collected in the subset "Exotic".

For the subset "Actual" that collects the most widely met types of distortions in the area of color image processing, the first two places are occupied by our metrics PSNR-HVS and PSNR-HVS-M. Their correlations with MOS are not ideal but they are, at least, larger than for other metrics. This allows recommending them for evaluating efficiency of image filtering and lossy compression. Matlab code for the metric PSNR-HVS-M is available from [40].

VII. Access To Tid2008, Conclusions And Acknowledgements

The archive TID2008 is available for downloading from [41]. This archive includes image files, the file containing the MOS values, the program for calculation of Spearman and Kendall correlations, the readme file where it is explained how to exploit the database. Also, archive contains the values of most known quality metrics calculated for TID2008. TID2008 occupies about 1 GB on a hard disk and about 600 MB in the archive.

We plan to regularly update the versions of this database. In particular, updated versions will provide more reliable data (in statistical sense) due to taking into account the results of future experiments. Moreover, new versions will include new types of distortion that take place in different applications of image processing and/or those distortions that might correspond to new peculiarities of HVS found in future experiments.

Finally, we would like to stress again to the following advantages of TID2008. First, it satisfies main requirements to such databases and contains many different types of distortion that relate to various peculiarities of HVS. This is important since for any quality measure when it is desirable to check its correspondence to more features of HVS. Currently it is not clear in which situations and applications of image processing that might appear in future a metric will be used and what features of HVS will be of a major value. Note that our database allows determining drawbacks of metrics. If all metrics would produce good results for a given database, it could be due to the database too simple. Note, that our database TID2008 has demonstrated serious drawbacks of known quality metrics.

The authors would like to thank all the people in Finland, Ukraine, and Italy who assisted in the experiments performance. Special thanks go to Dr. Heikki Huttunen (Tampere, Finland) and Mikhail Zriakhov (Kharkov, Ukraine).

References

[1] A. Bovik, "Handbook of Image and Video Processing", Academic Press, ISBN-10: 0121197905, 2000.

[2] K. Barner and G. Arce, , "Nonlinear Signal and Image Processing: Theory, Methods, and Applications (Electrical Engineering & Applied Signal Processing Series)", CRC Press, ISBN-10: 0849314275, 2003.

[3] M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," Signal Processing, vol. 70, pp. 247-78, 1998.

[4] A. B. Watson, "Digital Images and Human Vision", MIT Press, ch.3, pp. 139-140, 1993.

[5] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian and J. Astola, "Locally adaptive image filtering based on learning with clustering", Proc. of Image Processing: Algorithms and Systems IV, pp. 94-105, 2005.

[6] 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.

[7] N. Ponomarenko, F. Silvestri, K. Egiazarian, Carli M., Astola J., Lukin V. "On between-coefficient contrast masking of DCT basis functions", Proc. of the Third International Workshop on Video Processing and Quality Metrics. - USA, 2007. - 4 p.

[8] Z. Wang, E. P. Simoncelli, and A. C. Bovik "Multi-scale structural similarity for image quality assessment", IEEE Asilomar Conference on Signals, Systems and Computers, pp. 1398-1402, 2003.

[9] D.M. Chandler, and S. S. Hemami. "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images", IEEE Transactions on Image Processing, vol. 16 (9), pp. 2284-2298, 2007.

[10] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms", IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 3441-3452, 2006.

[11] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, and F. Battisti "Color Image Database for Evaluation of Image Quality Metrics", Proc. of the International Workshop on Multimedia Signal Processing, pp. 403-408, 2008.

[12] N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, and V. Lukin, "Metrics performance comparison for color image database", Proc. of the 4th International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2009, 6 p.

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

[14] Kodak Lossless True Color Image Suite: http://r0k.us/graphics/kodak/

[15] R.S. Berns, "Principles of Color Technology", John Wiley & Sons, ISBN 0-471-19459-X, 2000.

[16] A.V. Bazhyna, K.O. Egiazarian, N.N. Ponomarenko, and V. Lukin, "Compression of noisy Bayer pattern color filter array images", Proc. of the SPIE Conference on Computational Imaging V, vol. 6498, January, 2007.

[17] B.Bayer, "Color imaging array", U.S. Patent 3971065, 1976.

[18] A. Papoulis, "Poisson Process and Shot Noise", *Ch. 16 in Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, pp. 554-576, 1984.

[19] A. Foi, S. Alenius, M. Trimeche, V. Katkovnik, and K. Egiazarian, "A spatially adaptive Poissonian image deblurring", Proc. of IEEE International Conference on Image Processing,pp. 925-928, 2005.

[20] X. H. Han, Y. W. Chen, Z. Nakao, and H. Lu, "ICA-domain filtering of Poisson noise images", Proc. of the SPIE Third International Symposium on Multispectral Image Processing and Pattern Recognition, vol. 5286, pp. 811-814, 2003.

[21] I.J. Cox, M.L. Miller, J.A. Bloom, J. Fridrich, and T. Kalker, "Digital Watermarking and Steganography. 2nd Edition", Morgan Kaufmann, ISBN-13: 978-0-12-372585-1, 2008.

[22] M. Carli, "Perceptual Aspects in Data Hiding", Thesis for the degree of Doctor of Technology, Tampere University of Technology, 2008.

[23] K.N. Plataniotis and A.N. Venetsanopoulos, "Color Image Processing and Applications", Springer Verlag, ISBN 3-540-66953-1, 2000.

[24] E. Vansteenkiste, D. Van der Weken, W. Philips, and E. Kerre, "Psycho-visual quality assessment of state-of-the-art denoising schemes", Proceedings of EISIPCO, 2006, 5 p.

[25] O.G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising-part I: theory", IEEE Transaction on Image Processing, vol. 15, issue 3, pp. 555-571, 2006.

[26] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering", IEEE Transactions On Image Processing, vol. 16, issue 8, pp. 2080-2095,2007.

[27] "ITU (2002). Methodology for Subjective Assessment of the Quality of Television Pictures Recommendation BT.500-11", International Telecommunication Union, Geneva, Switzerland.

[28] K. Egiazarian, J. Astola, N. Ponomarenko, and V. Lukin, F. Battisti, M. Carli, "New full-reference quality metrics based on HVS",Proc. of the Second International Workshop on Video Processing and Quality Metrics, Scottsdale,, 2006, 4 p.

[29] B. W. Keelan and H. Urabe, "ISO 20462, A psychophysical image quality measurement standard", Image Quality and System Performance, Proc.of SPIE-IS&T Electronic Imaging, vol. 5294, pp. 181 -189, 2004.

[30] M. Gaubatz "Metrix MUX Visual Quality Assessment Package": http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[31] H. R. Sheikh and A. C. Bovik, "Image Information and Visual Quality", IEEE Transactions on Image Processing, vol. 15, pp. 430-444, 2006.

[32] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, A. Bovik "Image Quality Assessment Based on a Degradation Model", IEEE Transanction on Image Processing, vol. 9, pp. 636-650, 2000.

[33] Z. Wang andA. C. Bovik, "A universal image quality index", IEEE Signal Processing Letters, vol. 9, pp. 81–84, 2002.

[34] B. W. Kolpatzik and C. A. Bouman, "Optimized Universal Color Palette Design for Error Diffusion", Journal Electronic Imaging, vol. 4, pp. 131-143, 1995.

[35] B. Kolpatzik and C. Bouman, "Optimized Error Diffusion for High Quality Image Display", Journal Electronic Imaging, vol. 1, pp. 277-292, 1992.

[36] H. R. Sheikh, A.C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics", IEEE Transactions on Image Processing, vol.14, no.12, pp. 2117-2128, 2005.

[37] T. Mitsa, and K. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms", IEEE International Conference on Acoustic, Speech, and Signal processing, , vol. 5, pp. 301-304, 1993.

[38] A. B. Watson, “DCTune: A technique for visual optimization of DCT quantization matrices for individual images,” *Soc. Inf. Display Dig. Tech. Papers*, vol. XXIV, pp. 946–949, 1993.

[39] F. De Simone, D. Ticca, F. Dufaux, M. Ansorge, and T. Ebrahimi, “A Comparative study of color image compression standards using perceptually driven quality metrics”, Proc.of SPIE, Applications of Digital Image Processing XXXI, vol. 7073, pp. 70730Z-70730Z-11, 2008.

[40] PSNR-HVS-M page: http://ponomarenko.info/psnrhvsm.htm

[41] TID2008 page: http://www.ponomarenko.info/tid2008.htm