Signal Processing: *Image Communication* 30 (2015) 57–77

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Image database TID2013: Peculiarities, results and perspectives

Nikolay Ponomarenko [a], Lina Jin [b], Oleg Ieremeiev [a], Vladimir Lukin [a], Karen Egiazarian [b,*], Jaakko Astola [b], Benoit Vozel [c], Kacem Chehdi [c], Marco Carli [d], Federica Battisti [d], C.-C. Jay Kuo [e]

[a] *National Aerospace University, Dept of Transmitters, Receivers and Signal Processing, 17 Chkalova St, 61070 Kharkov, Ukraine*
[b] *Tampere University of Technology, Institute of Signal Processing, PO Box-553, FIN-33101 Tampere, Finland*
[c] *University of Rennes 1—IETR, CS 80518, 22305 Lannion Cedex, France*
[d] *University of Rome III, Via Ostiense, 161 Rome, Italy*
[e] *Media Communications Lab, USC Viterbi School of Engineering, SAL 300, Los Angeles, CA, USA*

ARTICLE INFO

Article history:
Received 3 April 2014
Received in revised form
22 October 2014
Accepted 22 October 2014
Available online 4 November 2014

Keywords:
Image visual quality metrics
Image denoising
Image lossy compression

ABSTRACT

This paper describes a recently created image database, TID2013, intended for evaluation of full-reference visual quality assessment metrics. With respect to TID2008, the new database contains a larger number (3000) of test images obtained from 25 reference images, 24 types of distortions for each reference image, and 5 levels for each type of distortion. Motivations for introducing 7 new types of distortions and one additional level of distortions are given; examples of distorted images are presented. Mean opinion scores (MOS) for the new database have been collected by performing 985 subjective experiments with volunteers (observers) from five countries (Finland, France, Italy, Ukraine, and USA). The availability of MOS allows the use of the designed database as a fundamental tool for assessing the effectiveness of visual quality. Furthermore, existing visual quality metrics have been tested with the proposed database and the collected results have been analyzed using rank order correlation coefficients between MOS and considered metrics. These correlation indices have been obtained both considering the full set of distorted images and specific image subsets, for highlighting advantages and drawbacks of existing, state of the art, quality metrics. Approaches to thorough performance analysis for a given metric are presented to detect practical situations or distortion types for which this metric is not adequate enough to human perception. The created image database and the collected MOS values are freely available for downloading and utilization for scientific purposes.

1. Introduction

Digital images have become an important part of everyday life and their quality is of primary importance for numerous applications [1–3]. Since humans are the main consumers of this type of information, it is important to address the problem of understanding the visual quality of digital images. To provide customers with high quality

* Corresponding author.
E-mail addresses: nikolay@ponomarenko.info (N. Ponomarenko), lina.jin@tut.fi (L. Jin), ol.eremeev@gmail.com (O. Ieremeiev), lukin@ai.kharkov.com (V. Lukin), karen.egiazarian@tut.fi (K. Egiazarian), jaakko.astola@tut.fi (J. Astola), benoit.vozel@univ-rennes1.fr (B. Vozel), kacem.chehdi@univ-rennes1.fr (K. Chehdi), marco.carli@uniroma3.it (M. Carli), federica.battisti@uniroma3.it (F. Battisti), cckuo@sipi.usc.edu (C.-C. Jay Kuo).

http://dx.doi.org/10.1016/j.image.2014.10.009
0923-5965/

of service, image quality metrics adequate to Human Visual System (HVS) are needed. Many visual quality metrics have been exploited in applications such as image and video lossy compression, image denoising, quality control of image printing and scanning, remote sensing, watermarking [2–8], etc.

There are several unsolved problems in design, verification, and use of visual quality metrics, as demonstrated in [2–5]. One of the problems is that HVS is very complex and it is difficult to both study and model HVS. For this reason, existing metrics attempt to incorporate a limited number of HVS specific features trying to match HVS judgment as close as possible. The evaluation of such matching (correspondence) is a challenging task as well.

One difficulty consists in the fact that performing subjective tests for collecting the mean opinion scores (MOS) is expensive and time consuming. Many observers are required for assessing the visual quality of a quite large number of distorted images collected in the so-called image databases. Then, mean opinion scores of image quality are determined [9,10] and further exploited. Note that different methodologies for database creation and carrying out experiments have been proposed. There are also different scales for MOS, different ways to remove outliers (abnormal experiments), and various recommendations to observers [5,9–11]. Because of this, the availability of databases of distorted images is quite reduced. Chandler, in [5], mentioned 20 existing databases which can be used for assessment of visual quality metrics. Among them, most used are, probably, LIVE [10], TID2008 [11], and Toyama [12]. Good databases CSIQ and IVPL [13,14] have appeared recently and design of new databases continues [5].

Available databases need to be improved. There are, at least, four reasons behind this need. A first reason is that intensive use of a given database reveals its drawbacks and limitations. These can be, for example, a limited number of distortions types. This was the motivation for creating TID2008 database, containing 17 types of distortions to extend the 5 types of distortions present in the LIVE database. A second reason is in the technology evolution pushing new consumer electronic devices and new applications that are characterized by new types of distortions or combinations of distortion types. To adequately cope with these distortions, test sets containing the new impairments should be designed and their perceptual impact addressed. A third reason is that creation of a new database leads to a certain “competition” among researchers. MSSIM has been reported as the best metric applied to TID2008 in [11] but quite many new metrics have overcome this result later, in the period 2009–2013. Such a competition is positive for both theory and practice since it results in more universal visual quality metrics or, at least, in designing metrics well suited for certain sets of distortion types. A fourth reason is in the suggestion we received to increase the number of color distortions and the just noticeable distortions present in TID2008.

Based on these considerations, a novel image database, TID2013, has been designed and published [15,16]. Due to limited space in Conference Proceedings, many important aspects have not been addressed and described in [15,16]. In particular, this relates to motivations for selecting new types of distortions for the new database and methodologies of their simulation. Besides, a limited set of visual quality metrics has been tested for TID2013. However, the most important aspects, to our opinion, concern the results already obtained for the database TID2013 and perspectives of its exploitation in future.

The rest of the paper is organized as follows. Requirements to distorted image databases and peculiarities of the database TID2013 that differ it from TID2008 are considered in Section 2. Section 3 pays a special attention to the description and motivation of the new types of distortions. Section 4 describes the methodology used in the subjective experiments and the collected results. Analysis of results obtained for a set of popular visual quality metrics is given in Section 5. Section 6 concerns a special analysis for some known metrics showing how to determine drawbacks of metrics exploiting TID2013. Finally, Section 7 describes the modalities for accessing to the database and gives information that can be useful for people planning to employ TID2013.

2. Peculiarities of the new database

As it has been mentioned above, the database TID2008 is the predecessor of TID2013. After its creation five years ago, TID2008 served several purposes, both main and auxiliary. The main purpose of TID2008 was the analysis and verification of full-reference metrics [15,17]. To this aim, it has been extensively used [18–22]. Meanwhile, TID2008 has been also used for auxiliary purposes as testing and efficiency analysis of blind methods for noise variance estimation [23,24], colour image denoising techniques [25], and verification of no-reference metrics [26].

There are basic requirements to databases intended for HVS metric design and assessment. Such a database *should contain a reasonably large number of etalon colour images of various content*. TID2008 contains 25 reference (distortion-free, etalon) colour images 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 —see all 25 distortion-free images as shown in Fig. 1. As it can be seen, the test images are of different content, some of them are quite textural ones whilst others contain large quasi-homogeneous regions. Thus, the abovementioned requirements are satisfied.

Size of images in a database can be, in general, debated. However, there are some restrictions and recommendations. First, restrictions deal with methodology of experiment carrying out. Two or three images are usually displayed simultaneously at the monitor screen and their quality is to be assessed (compared). This means that image size has to allow simultaneous full representation of these images at screen of devices used in experiments. For both TID2008 and TID2013 it was supposed that images were displayed at computer monitors. Because of this, all images were of the same fixed size 512 × 384 pixel in TID2008 and we have kept the same size for the images in TID2013. There are two things to be considered. First, image size might influence image perception and this

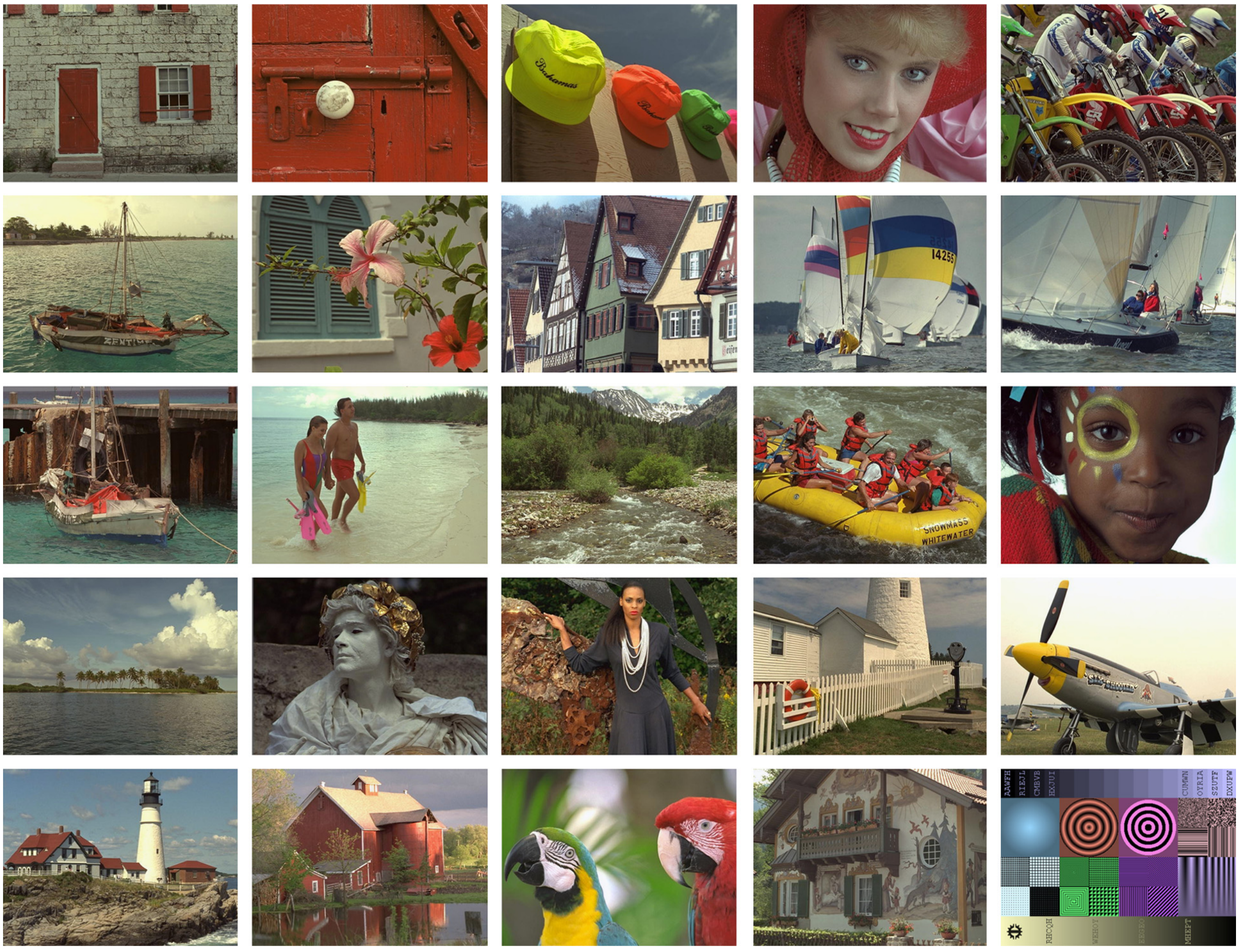


Fig. 1. Reference images in databases TID2008 and TID2013.

motivates using the same size. Second, some participants of the conferences EUVIP2013 and ACIVS2013 have expressed a desire to have larger size test images being interested in such modern applications as HDTV. We are not able to satisfy them but the need in a database for high-resolution applications should be addressed somehow.

Another requirement to a database is that *it should contain image distortions typical for practice* and, simultaneously, these *distortions have to relate to certain peculiarities of HVS*. Almost everybody has nowadays met with distortions due to conditions of image acquiring—noise and blur, chromatic aberrations. Many people encountered distortions originating due to compression and data transmission errors. There are also distortions due to specific operations of image processing as denoising, mean and contrast changing, etc. Seventeen types of distortions were taken into account while creating TID2008. They are presented in 17 first rows in Table 1 with explanations what are the main applications a given distortion can be met with and what peculiarity of HVS this distortion type relates to. More details concerning the reasons for including these distortion types into TID2008 and their modeling can be found in [17].

The main difference of TID2013 compared to TID2008 is that the new database includes 7 new types of distortions marked by items from 18 to 24 in Table 1. These types of distortions will be considered more in detail in Section 3. Here, we highlight two main aspects. First, we have tried to pay more attention to “color” distortions (#18, 22, and 23) since the percentage of grayscale images that are in use today decreases and preservation of color information becomes more and more important. Second, we attempted to consider new applications (distortion types #19…22 and 24) for which images with the corresponding distortions are absent in existing databases and for which adequate visual quality metrics are expected to be of prime importance.

Table 1 needs some additional comments. By “eccentricity” (see the right-hand column) of HVS we mean noticeability of distorted fragments in images, their difference compared to surrounding fragments of reference image in texture, color or other features. “Robustness” relates to human ability to “filter out” noise (including impulse noise) in images with “restoring” them. “Evenness of distortions” means that humans perceive distortions spread uniformly and distortions placed compactly in different manner. Also note that a certain type of

Table 1
Types of distortions used in image databases TID2008 and TID2013 and their correspondence to practice and HVS.

№	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	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
18	Change of color saturation	Image compression, Image acquisition	Color sensitivity
19	Multiplicative Gaussian noise	Image acquisition, image denoising	Adaptivity, robustness
20	Comfort noise	Image compression	Eccentricity
21	Lossy compression of noisy images	Image compression, image denoising	Spatial frequency sensitivity, local contrast sensitivity
22	Image color quantization with dither	Image registration	Color sensitivity, local contrast, spatial frequency
23	Chromatic aberrations	Image acquisition	Color sensitivity, local contrast sensitivity
24	Sparse sampling and reconstruction	Image compression, image reconstruction	Spatial frequency sensitivity, local contrast sensitivity

distortions may relate to a mentioned application only partly. Meanwhile, in practice, there are usually several types of distortions typical for each given application.

One more requirement to an image database is that images in the database should be challenging for visual quality assessment. This requirement means, in the first order, that number of situations when all quality metrics evidence in favor of a given image among two compared should not be large. Let us give examples of such situations. For instance, it might happen that an observer during an experiment will be asked to compare visual quality of images presented in Fig. 2.

Then, the decision is clear and fast since the right-hand image obviously has better visual quality and visual quality metrics also confirm this. Fig. 3 presents another undesirable type of comparison (quality assessment) situation when the same type but different levels of distortions distort two images presented at a monitor screen. The results of comparisons are clear and predictable. HVS-metrics usually have perfect correspondence to such images. This may cause an illusion that metrics perform well and there are no problems with an adequate assessment.

An example in Fig. 3 shows that, on one hand, there should not be a large number of distortion levels. Four or five levels are usually enough for a database [10,11] and there were four levels of distortions in TID2008. On the other hand, the database TID2008 has been criticized for not having images with almost invisible (not apparent) distortions [27]. To cope with this task, we have introduced the fifth distortion level for all test images and all distortion types present in TID2013. This added level approximately corresponds to a peak signal-to-noise ratio (PSNR) equal to 33 dB (recall that for images with other four levels of distortions in TID2008 and TID2013 the PSNR values are approximately equal to 30, 27, 24, and 21 dB).

The presence of five levels of distortions is one more distinctive difference of the new database TID2013 compared to TID2008. As a result, TID2013 contains 3000 distorted images (25 test images with 24 types and 5 levels of distortions) in opposite to 1700 distorted images in TID2008. There are also some differences but they relate not directly to the database but to a methodology to obtain MOS and conditions to carry out experiments. These differences will be discussed in the next section.

3. New types of distortions

During the design phase of TID2013, we had to decide how many new types of distortions have to be exemplified

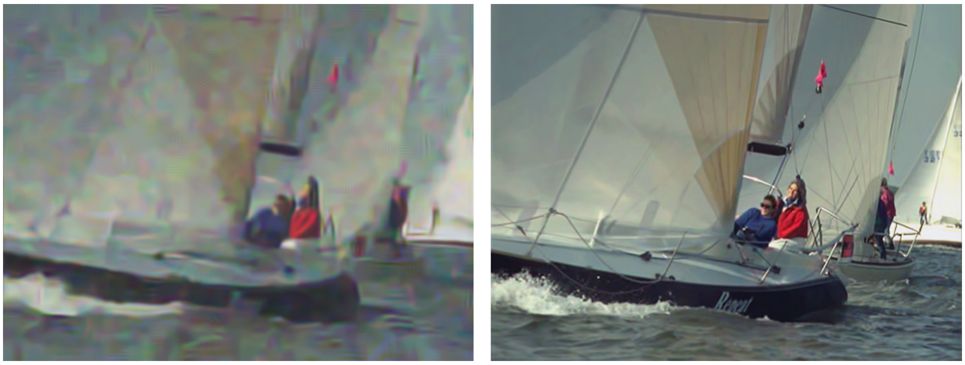


Fig. 2. Example of undesired practical situation in pair-wise comparisons of visual quality of two distorted images.

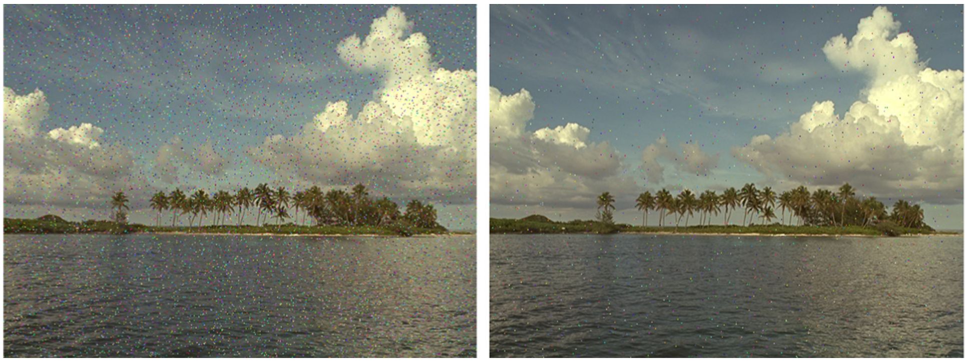


Fig. 3. Example of undesired practical situation in pair-wise comparisons of visual quality of two distorted images.

in a new database and what should be these types. Certainly, it was necessary to add such types of distortions that are valuable from both theoretical and practical points of view. We created a list of possible types of distortions that had more than ten positions. This list was discussed in teams of five countries the authors of this paper are from. Several factors were taken into account as does a new type of distortions considerably differ from the ones already existing in TID2008, how often customers and industry deal with a given type of distortion, has a given new technology perspectives in future, etc.

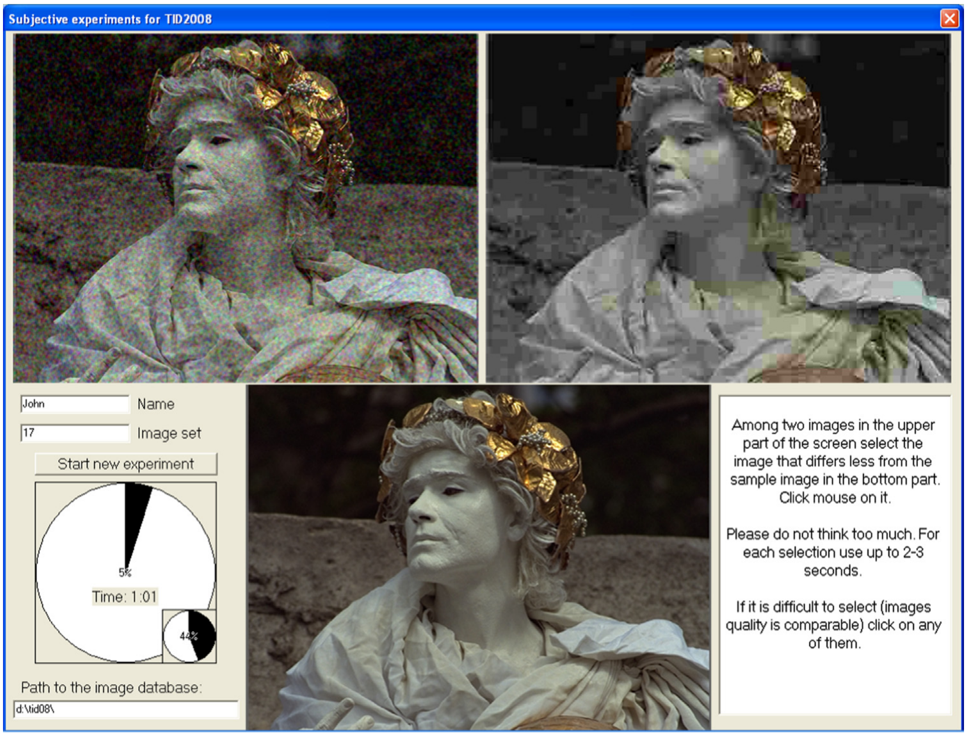
One could ask why not to add more types of distortions? The answer is the following. Creating the database, we had to take into account some limitations. First, adding more types of distortions leads to a larger number of distorted images for a given reference image, resulting in a greater time spent for each experiment. Meanwhile, this time should not be too large to prevent observer's tiredness. Second, we need to have an even number of distorted images for each reference image. Then it is possible to make each distorted image to participate in equal number of comparisons.

These limitations can be still unclear without a brief description of the methodology of experiments. At a monitor, a pair of distorted images (in the upper part) and the corresponding reference image (in the lower part) are simultaneously displayed (see an example in Fig. 4). An observer was asked to choose a better distorted image (between two upper ones). By a “better” image we mean the image that differs less from the reference one. After the first and each next selection, a given pair of distorted images disappears and two different (new) distorted images appear. Then, comparison (selection) is done again. Each distorted image participates in equal number of pair-wise comparisons (more in detail, nine comparisons, see next section).

Since five levels of distortions are used in TID2013, we need even number of distortion types to have even number of distorted images for each reference one.

Taking into consideration aforementioned peculiarities and limitation, we have decided to introduce just seven new types of distortions. As a result, we have got the total number of distortion types equal to 24 and a total number of 120 distorted images for each reference.

Let's consider each new type of distortions in more details. A change in *color saturation* (distortion type # 18, see example in Fig. 5) may come as a result of different factors at the stages of image acquisition and processing. In particular, it can arise due to a large quantization of colour components in JPEG-based compression of images and video [28]. Such a distortion might also take place in colour image printing. All simulations have been carried out in Matlab. Modelling of these distortions has been performed after image transformation from RGB to YCbCr colour space using function rgb2ycbcr. The component *Y* (intensity) remained untouched and the components Cb



Subjective experiments for TID2008

Name: John

Image set: 17

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. 4. Screenshot of the software used in experiments that illustrates positions of displayed images.

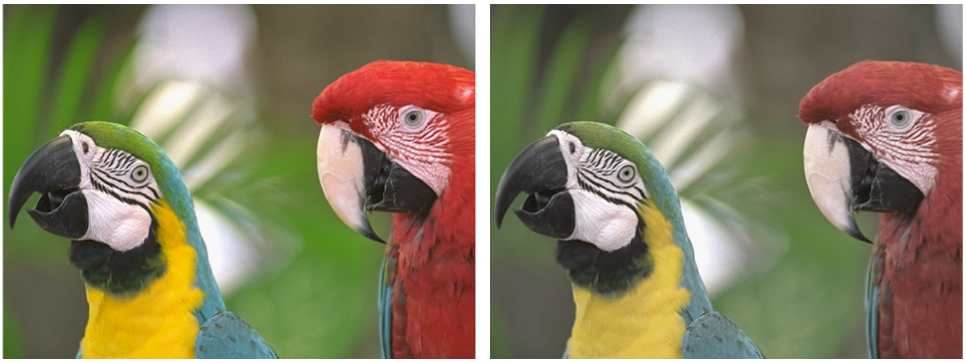


Fig. 5. Example of color saturation effect: distortion free (left) and distorted (right) images.

and Cr were transformed as $\text{Cb}=128+(\text{Cb}-128)\times K$ and $\text{Cr}=128+(\text{Cr}-128)\times K$ where K is a variable parameter. After such a transformation, the obtained image has been converted to the original color space using function ycbcr2rgb. K equals to 1 relates to the absence of distortion, the use of smaller values of K leads to making image less ‘colorful’. K equals to 0 makes a color image to look as a grayscale one. K values have been adjusted to provide a desired PSNR. In some cases, to provide low PSNRs (21 or 24 dB) we needed to use negative values of K resulted in the specific (inverse) color distortions.

Multiplicative Gaussian noise (distortion type # 19, see example in Fig. 6) represents a wide class of signal-dependent noise. As far as we know, such type of distortions is absent in other databases. Meanwhile, signal-dependent noise occurs in images in many applications where visual quality of images is of a prime importance [29] including single- and multichannel radar imaging [30], multispectral remote sensing, medical imaging [31], etc. A multiplicative Gaussian noise has been simulated separately (independently) for each colour (RGB) component with equal variance of multiplicative noise σ_μ^2 in all components. The values of σ_μ^2 have been adjusted individually for each reference image and each distortion level to provide required values of PSNR.

Comfort noise (distortion type #20) is a specific type of distortions. It is known that humans do not pay much attention to a realization of the noise present in a given image. Similarly, humans sometimes cannot distinguish realizations of texture if the texture fragments have the same parameters. These properties are already exploited in lossy compression of images and video [32–34] to

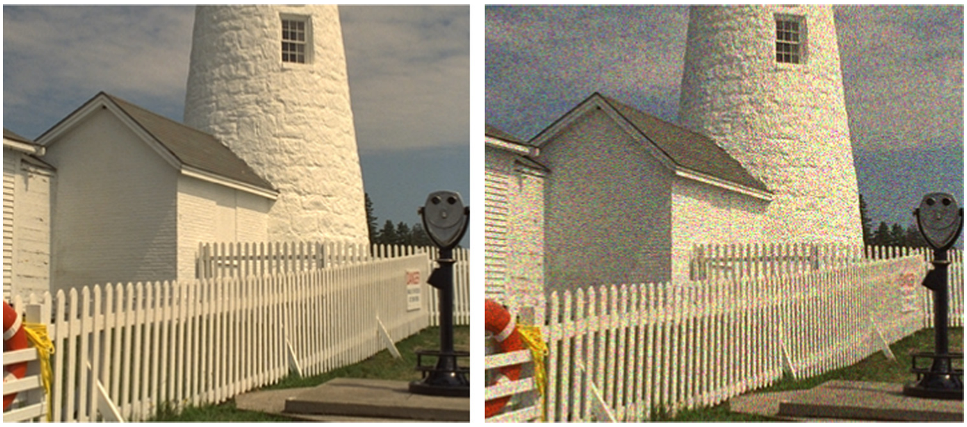


Fig. 6. Fragments of the same test image corrupted by multiplicative noise of different level.

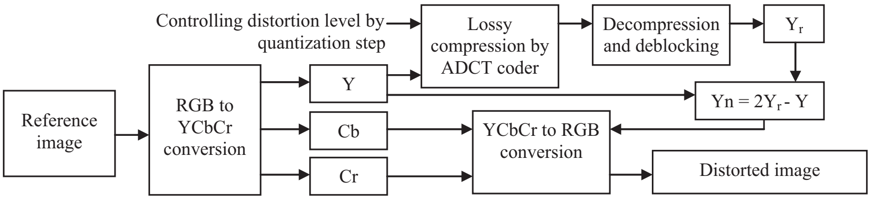

Controlling distortion level by quantization step
Lossy compression by ADCT coder
Decompression and deblocking
Y_r
Reference image
RGB to YCbCr conversion
Y
Cb
Cr
Yn = 2Y$_r$ - Y
YCbCr to RGB conversion
Distorted image

Fig. 7. Modeling of comfort noise distortion.

simultaneously attain larger compression ratio and natural appearance of decompressed data. Such methodology leads to a quite large difference between reference and distorted images in terms of standard metrics such as MSE or PSNR whilst visually these images might look very similar. Comfort noise distortions have been modelled as follows (see Fig.7).

An original image was converted from RGB color space to YCbCr. Then, a comfort noise was modeled separately for each components Y, Cb, and Cr. Let us explain a modeling procedure on example of the luminance (Y) component. The image Y is lossy compressed by the coder ADCT [35]. Then, the image is decompressed and post-processed for blocking artifact removal. As a result, the reconstructed image Y_r is obtained. It is supposed that losses introduced by the compression are mainly related to a noise. Then a noisy part of an original image can be roughly estimated as $Y_n = Y - Y_r$. Then, a comfort noise is simulated as $Y_r - Y_n$, i.e., noise with inverse "polarity" is added to the image. Thus, the image distorted by a comfort noise Y_d is modelled as $Y_d = Y_r + Y_r - Y$. Similar procedures are applied also to color components Cb and Cr. Then, an inverse color space conversion is performed using the function ycbcr2rgb. A desired PSNR is reached individually by varying a compression ratio (CR) for ADCTC (in fact, CR is controlled by a quantization step for this coder). Unfortunately, most reference images in TID2013 do not contain contrast noise-like textures. Because of this, a comfort noise (as we understand it) has been provided only for low levels of distortions (PSNR approximately equal to 33 and 30 dB). For larger distortion levels, information content of images occurs to be distorted as well.

Examples of comfort noise in images are shown in Fig. 8.

The next new type of distortion is *lossy compression* of noisy images (# 21). Such type of distortions takes place in compressing both video and images acquired in non-perfect conditions [33,34] making it very important in practice. Besides, it has been stated by many researchers that usually there are several types of distortions simultaneously present in images and video whilst databases commonly contain images with "pure" distortions. To make an impact of noise and lossy compression comparable, the distortions have been modelled as follows (see Fig. 9).

Additive white Gaussian noise with variance σ^2 has been added to each colour component where noise was independent for colour components. After this, lossy compression has been performed using DCT-based coder ADCT [35] with the quantization step set equal to 1.73σ. Such quantization step is chosen in order to provide a visibility of both distortions from compression noise and residual noise. Noise standard deviation has been individually adjusted for each test image to provide a desired value of PSNR. Examples of images with the considered type of distortions are presented in Fig. 10. As it can be seen, the distortions can be quite specific.

Image colour quantization with dither (# 22) is typical in image printing. It is one more popular application which is paid particular attention nowadays [36]. Distortions of this type have been modelled using the Matlab function rgb2ind. It converts an RGB image to the indexed image using dither. To provide a desired PSNR, the number of quantization levels was adjusted individually for each test

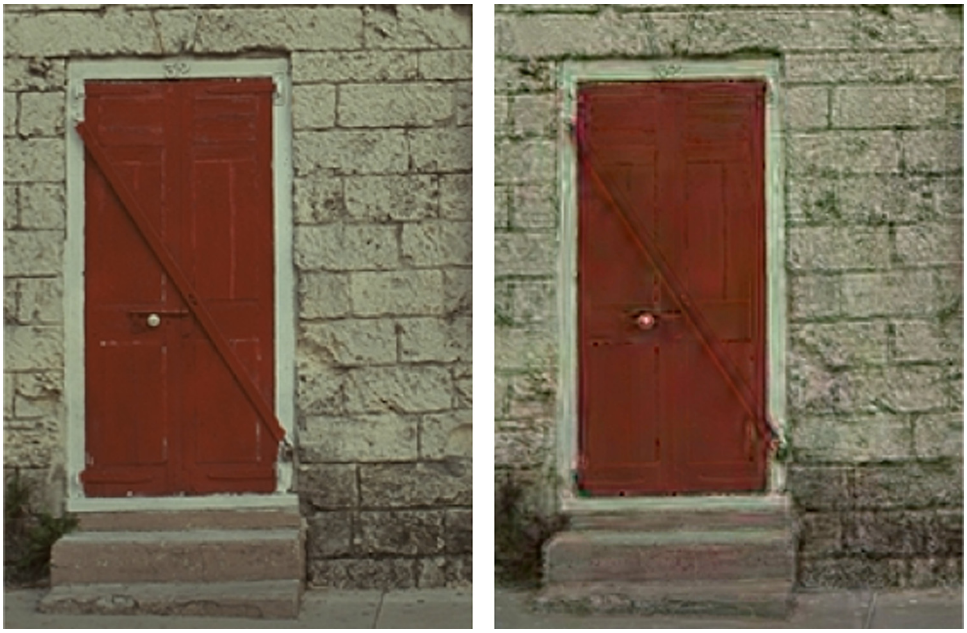


Fig. 8. Fragments of the same test image corrupted by comfort noise of different levels.

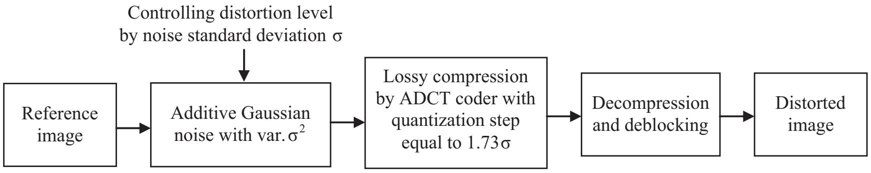

Controlling distortion level by noise standard deviation σ
Reference image
Additive Gaussian noise with var. σ²
Lossy compression by ADCT coder with quantization step equal to 1.73σ
Decompression and deblocking
Distorted image

Fig. 9. Modeling of distortions of noisy image lossy compression.

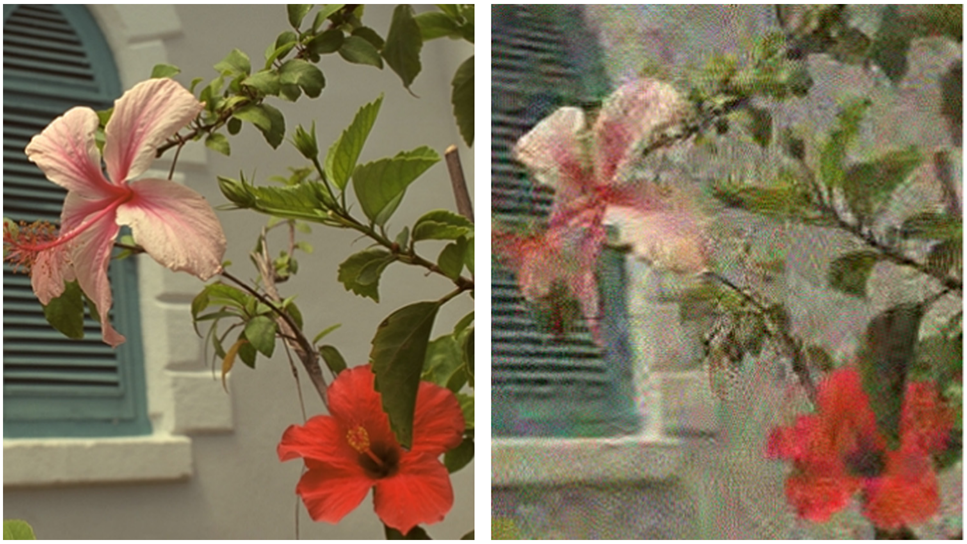


Fig. 10. Fragments of distortion-free (left) and distorted (right) image corrupted by additive noise and lossy compression.

image. Examples of images with this type of distortions are shown in Fig. 11.

Chromatic aberrations (distortion type #23) might take place at image acquisition stage but similar effects can also appear at stages of image transformations. It is quite annoying type of distortions especially in places of high contrasts and if a distortion level is high. Chromatic aberrations have been modelled by carrying out mutual shifting of *R*, *G*, and *B* components with respect to each other (see Fig. 12).

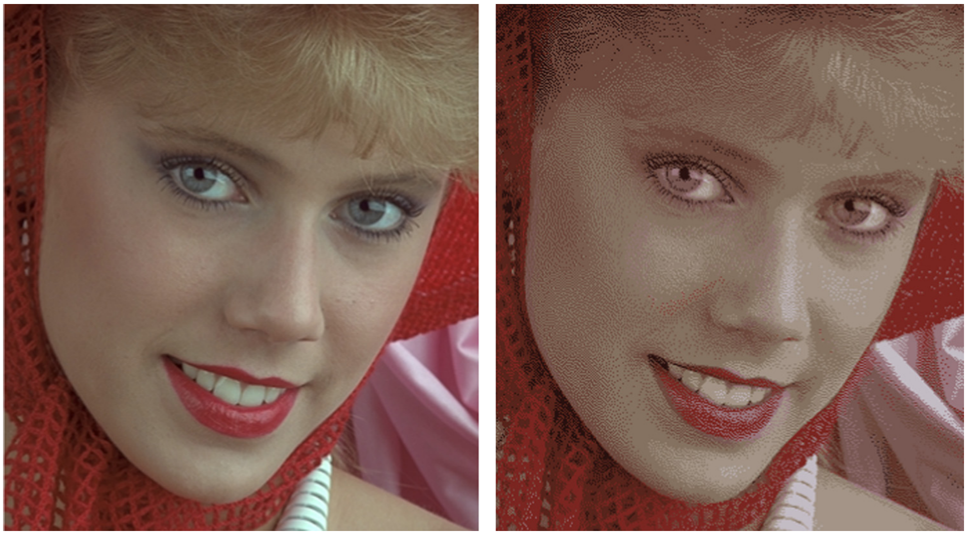


Fig. 11. Fragments of distortion-free (left) and distorted (right) image with dither.

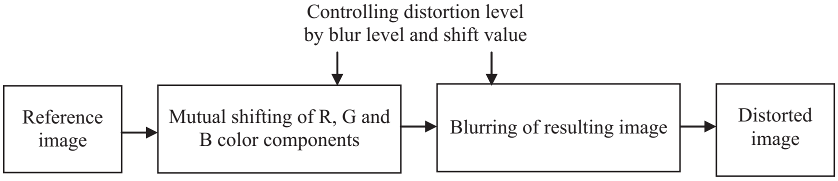

Controlling distortion level
by blur level and shift value
Reference image
Mutual shifting of R, G and B color components
Blurring of resulting image
Distorted image

Fig. 12. Modeling of distortions of chromatic aberrations.

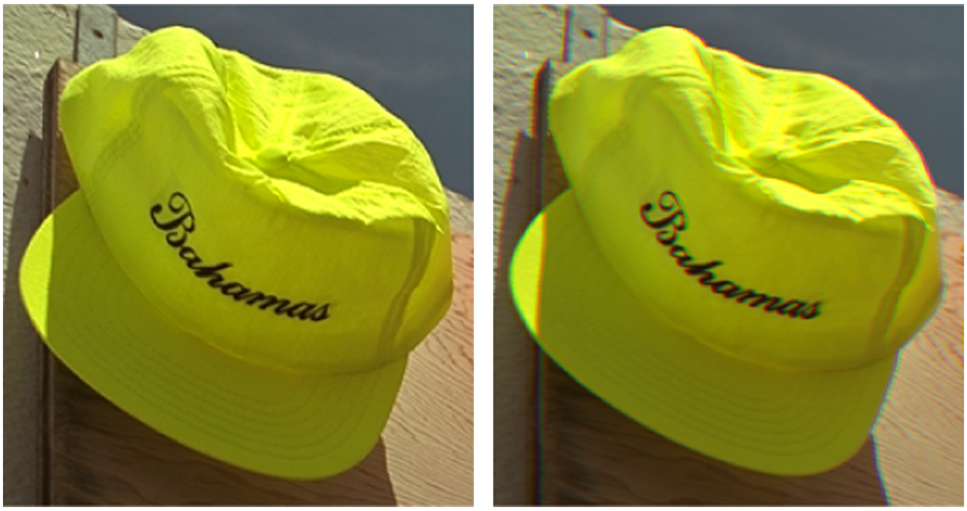


Fig. 13. Fragments of distortion-free (left) and distorted (right) image with chromatic aberrations.

Besides, further slight blurring of shifted components has been performed. Shifting and blurring parameters have been adjusted to provide a desired PSNR. An example is shown in Fig. 13.

Finally, the last distortion type (#24) relates to compressive sensing (sparse sampling and reconstruction) that has become a hot research topic [37,38]. As far as we know, HVS-metrics have not been exploited in this area yet although their usefulness is expected. An example of distortions for this application is presented in Fig. 14 though they can depend upon a method of compressive sensing used.

For us, it was convenient to use the method [38] and available software for obtaining reconstructed images with distortions. As earlier, modelling is carried out separately for components Y, Cb, and Cr. Some details for Y component are explained by Fig. 15.

The Y component image is subject to the 2D discrete cosine transform (DCT) applied to the entire image getting

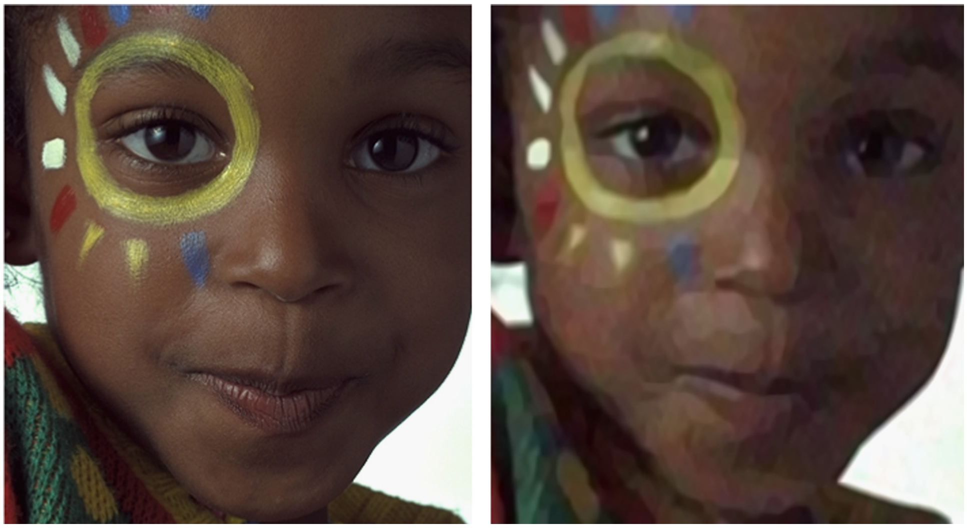


Fig. 14. Fragments of distortion-free (left) and distorted (right) image obtained by compressive sensing.

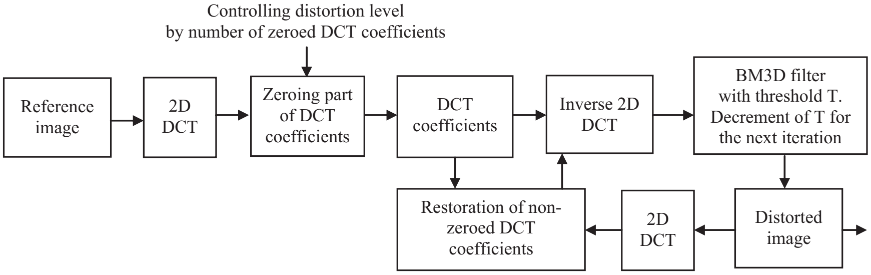

Controlling distortion level
by number of zeroed DCT coefficients
Reference image
2D DCT
Zeroing part of DCT coefficients
DCT coefficients
Inverse 2D DCT
BM3D filter with threshold T. Decrement of T for the next iteration
Restoration of non-zeroed DCT coefficients
2D DCT
Distorted image

Fig. 15. Modeling of distortions of compressive sensing.

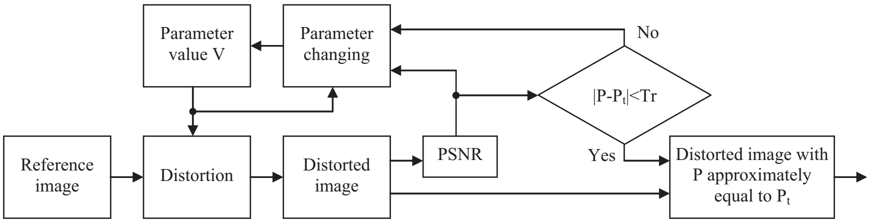


Fig. 16. Block diagram of obtaining distorted image with a required PSNR.

2D spectrum Y_{DCT} (matrix of DCT coefficients). A sufficient part of coefficients in Y_{DCT} is assigned zero values (larger number of zeroed coefficients leads to larger distortions). The matrix Y_{DCT} after zeroing some coefficient was saved as a matrix Y_{DCT0}. Then the following sequence of operations is performed in ten iterations for Y_{DCT}. Carry out inverse DCT: $Y_r = \mathrm{IDCT2}(Y_{DCT})$. Process Y_r by the BM3D filter [38] and obtain the filtered image Y_f. Apply DCT to it: $Y_{DCT} = \mathrm{DCT2}(Y_f)$. At the end of each iteration, correct those values in Y_{DCT} that are not equal to zero in Y_{DCT0}: $Y_{DCT}(Y_{DCT0} > 0) = Y_{DCT0}(Y_{DCT0} > 0)$. After the last iteration, the distorted image is obtained by inverse DCT: $Y_d = \mathrm{IDCT2}(Y_{DCT})$. A required PSNR is provided by varying a number of DCT components to which zero values are assigned.

As it follows from the description presented above, the new types of distortions introduced in the database TID2013 are quite different. It took a lot of time to carry out extensive computations in order to provide desired levels of distortions. These computations have been partly automated to simplify the process (see Fig. 16).

Desired values of PSNR for each reference image and distortion type and level have been provided by adjusting the corresponding parameter(s) of the simulation algorithm. Sometimes this required designing special iterative procedures with variable step of parameter changes as, e.g., parameter σ^2 for distortion type # 21. Procedures of parameter adjusting continued until a desired value PSNR (P_t) and obtained value PSNR (P) occur to differ less than a preset threshold (Tr). For some types of distortions this threshold was set equal to 0.025 dB, for more complex types of distortions it was 0.1 dB or even 0.2 dB.

At the very beginning of iterative procedures, we used two values of varied parameter V_1 and V_2 that surely provide considerably smaller PSNR (P_l) and essentially larger PSNR (P_h) than P_t. Then, by interpolation a value V between V_1 and V_2 was set with further making search interval narrower to converge to P_t.

4. Experiments description and results

4.1. MOS obtaining

Having a database, MOS is to be provided for each distorted image in it. There are several methodologies used to assess the visual quality of an image [39–41]. For example, the observers might be asked to assess the absolute quality of an image or its similarity to a reference one. Then, the subject judgment is expressed with a grading scale that can be of a different type. Five gradations have been used in [39] with the corresponding five categories as “Bad”, “Poor”, “Fair”, “Good”, and “Excellent”. A drawback of this methodology is that it might be difficult for an observer to assign gradations to the distorted images, especially at the beginning of experiments when an observer has a little experience. This leads to the observer's willingness to change the previously assigned grade. Because of this, the observers sometimes undergo a training phase where some examples of the distortions that will be met in tests are offered before just experiments.

When obtaining MOS for TID2013, we have used another methodology that has been previously employed for carrying out the subjective tests in TID2008. As it was mentioned above, three images have been displayed (tristimulus methodology, see Fig. 4) and an observer selects a better image between two distorted ones. Many experiment participants have accepted this methodology of comparisons as less annoying. To provide an accurate estimate of MOS, it is needed to carry out enough number of experiments and to remove those ones that have been distinguished as abnormal [41].

In more details, each observer in one experiment has carried out distorted image quality assessment for only one reference image. There are 120 distorted images (five levels and twenty four types of distortions) for each reference in TID2013. Each of 120 distorted images participated in nine pair-wise comparisons. The preferred image for each pair of displayed ones got one point. The winning points were summed-up to get the final score for each distorted image. “Competition” was organized in a manner similar to Swiss system in chess. After starting round which was absolutely random and a few other, pseudorandom rounds, “approximately the same strength players” (approximately the same visual quality images) competed in pairs.

Thus, each observer for one reference image had to carry out 540 pair-wise comparisons of visual quality. This took about 17 min on the average (recall that average time for one experiment in TID2008 was 13.5 min). According to the recommendations of ITU [41], the time for performing one experiment should not exceed 30 min to avoid tiredness and its influence on experiment outcomes. No one experiment carried out in laboratory conditions lasted more than 30 min. Therefore, ITU recommendations have been met.

Before starting the experiments, all observers were instructed. If an experiment was done in laboratory conditions, a tutor had passed instructions to experiment participants and the tutor was taking care over following the instructions during experiments. For experiments carried out via Internet, a participant had to read Instructions related to preferred (recommended) conditions and a methodology of experiments. In particular, it was recommended to use convenient (preferred) distance to a monitor and to compare visual quality of images for not more than a few seconds for each pair of distorted images.

Protocol of each experiment including results of pair-wise comparisons has been documented and saved. After getting the protocols from all observers, they were processed in a “robust” manner. By this we mean that abnormal results have been detected and rejected from further consideration. 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 [41] using the same methodology as in [39]. Note that abnormal results occurred with a probability about 2%. After this, the obtained results were averaged for each reference image. Thus, the obtained MOS has to vary from 0 to 9 and its larger values correspond to better visual quality.

It is interesting that in the resulting MOS there were no values equal to 0 or 9 (see MOS histogram in Fig. 17). Moreover, there were no MOS values larger than 7.5. This shows that conditions of comparisons were quite difficult especially for distorted images with rather high visual quality.

Experiments for TID2013 were conducted in five countries (Finland, France, Italy, Ukraine, USA). Three persons from other countries participated in experiments as well. This is because it was possible to carry out experiments both in laboratory conditions (under control of tutors) and

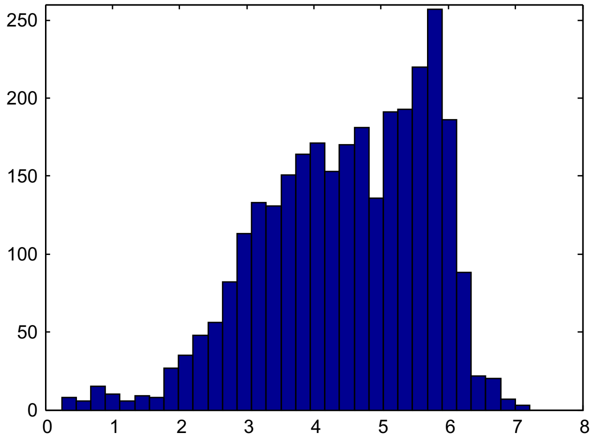


Fig. 17. MOS histogram for TID2013.

distantly via Internet. The results obtained in different countries were in good agreement (see details later).

Note that approximately equal number (over 30) of experiments was performed for each reference image. This allows stating that MOS is of approximately the same accuracy for all reference images.

Some other data describing experiments and accuracy of the estimated MOS are presented in Table 2. As it is seen, TID2013 is a leader in total number of experiments and number of elementary evaluations. Due to this, an accuracy of MOS estimation is practically the same as for TID2008.

Although experiment participants were instructed before starting experiments, subjective tests have been done in different conditions. In particular, different monitors were used, both LCD and CRT, mainly 19" and more with the resolution 1152 × 864 pixel. More than 300 observers have performed experiments via Internet. Most of participants were students although tutors and researchers also took part. Observation conditions varied in reasonable limits and we asked participants to use distance from monitors comfortable for them. All these do not correspond to stricter requirements imposed by ITU. However, in our opinion, visualization and analysis of image quality in slightly varying conditions provide reasonably good verification of quality metrics if these metrics are intended for visual quality assessment in practice in a priori unknown conditions. Non-identical conditions of experiments take into account how visual quality is assessed in everyday practice of computer users and Internet customers.

4.2. MOS property analysis

Let us consider some other properties of MOS. Its values for all 3000 distorted images in the database are presented as scatterplot in Fig. 18 where first (leftmost) 120 points correspond to the reference image #1, next 120 points relate to the distorted images that have the same reference image #2 and so on. This scatterplot shows that MOS values are most dense within the interval from 3 to 6 and, thus, the task of comparing image visual quality was not trivial.

MOS values averaged for all observers that carried out experiments are presented for each distortion type and level in Fig. 19. Each 5 level points for a given type of distortions (24 totally) are connected to see a tendency if it exists. For most types of distortions, the tendency is clear and obvious—average MOS decreases if distortion level becomes larger. The exceptions are Distortion types #15 and #17. Recall that distortion type #15 is Local block-wise distortions of different intensity where for level 1 one has a larger number of blocks than for other levels but contrasts of these blocks with respect to surrounding are smaller. The results in Fig. 19 show that for observers assessing visual quality it is more important what the total

Table 2
Comparison characteristics of Databases LIVE, TID2008 and TID2013.

N	Main characteristics	Test image database		
		LIVE database	TID2008	TID2013
1	Number of distorted images	779	1700	3000
2	Number of different types of distortions	5	17	24
3	Number of experiments carried out	161 (all USA)	Totally 838 (437—Ukraine, 251—Finland, 150—Italy)	Totally 971 (602—Ukraine, 116—Finland, 101—USA, 80—Italy, 72—France)
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	25,000	25,6428	524,340
6	Scale of obtained estimates of MOS	0..100 (stretched from the scale 1..5)	0..9	0..9
7	Variance of estimates of MOS	250	0.63	0.69
8	Normalized variance of estimates of MOS	0.083	0.031	0.035
9	Variance of MOS	–	0.019	0.018

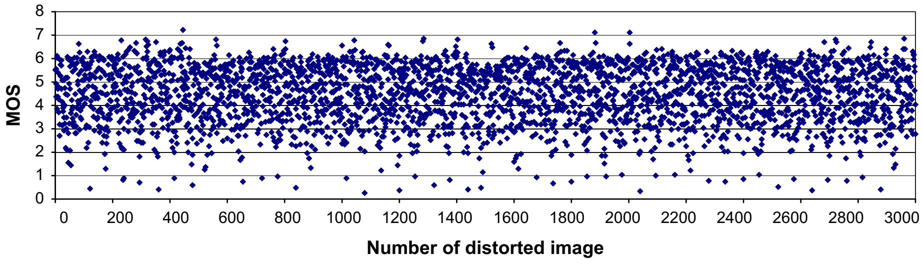


Fig. 18. Scatterplot of MOS for all 3000 distorted images in TID2013.

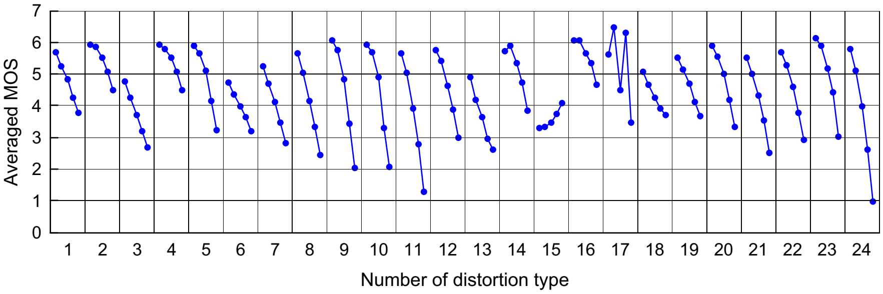


Fig. 19. Dependence of MOS on distortion type and level.

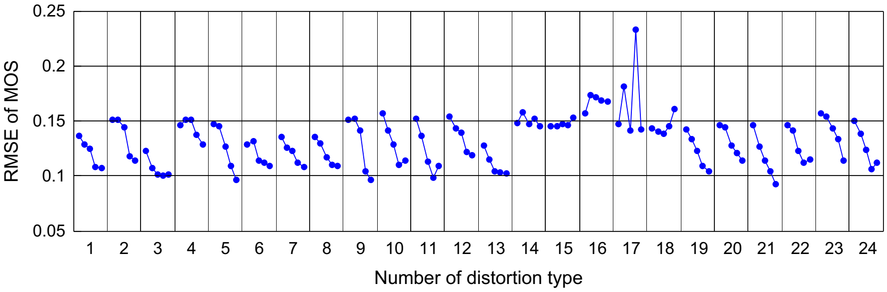


Fig. 20. Mean RMSEs of MOS for different types and levels of distortions.

area of such blocks (that “hide” useful information) is than what the block contrasts are. Distortion type # 17 relates to Contrast change (see Table 1). Level 1 corresponds to small contrast decreasing, level 2—to small contrast increasing, level 3—to a larger contrast decreasing, level 4—to a larger contrast increasing, level 5—to the largest contrast decreasing. The results in Fig. 19 clearly show that contrast increasing is perceived as better than contrast decreasing. However, there is certain “optimal” contrast change that approximately corresponds to level 1.

Except MOS values, it could be interesting to analyze deviation values (characterized by Mean RMSE) of MOS depending upon distortion type and level. The obtained results are presented in Fig. 20 where again for each type of distortion we have 5 points corresponding to five levels (starting from the leftmost point that relates to level 1). There is an interesting tendency here. RMSE values usually diminish if distortion level increases. This means that it was more difficult to undertake decisions in comparisons for distorted images of quite high visual quality than if one or two compared images were considerably distorted.

People had the smallest variations in judgments concerning the images with distortions #13 (JPEG2000 transmission errors) and # 3 (Spatially correlated noise). However, it was difficult for observers to assess the visual quality of images with distortion type #17 (Contrast change), especially for images with large contrast increase (level 4). Mean RMSE values are almost the same for all levels for distortion types #14, #15, #16, and #18. All these distortion types can be referred to the class (subset) called Exotic [42].

Note that the database TID2013 contains not only the file “mos.txt” of MOS values but also the file “mos_std.txt” of MOS standard deviations for each distorted image).

The obtained RMSEs of MOS allow analyzing what is the agreement between the results obtained in different countries and conditions—in laboratory and via Internet. For this purpose, we have calculated Spearman rank order correlation coefficient (SROCC) between mean opinion scores averaged for Ukrainian participants (602 experiments, all done in laboratory) and participants from other countries (369 totally, mostly carried out via Internet). The calculated SROCC value is equal to 0.955. We have also obtained SROCC for experiments performed on-line (139 experiments) and off-line (832 experiments). It is equal to 0.934.

To understand are these values relatively high or low, additional simulations have been done. We have simulated MOS with RMSE values obtained for the aforementioned groups (they vary from 0.15 for 832 experiments to 0.32 for 139 experiments). The obtained “ideal” SROCC values are equal to 0.989 and 0.984, respectively. This means that, on one hand, the correlation for experiments performed in different countries is high enough. Correlation for on-line and off-line experiments is also rather high. On the other hand, the correlation occurs to be influenced by conditions in which the experiments have been carried out and this influence is to be additionally studied.

Consider now Mean RMSE of MOS depending upon reference image and distortion level. The corresponding data are represented in Fig. 21.

Analysis shows that images that are more distorted are usually assessed more similarly by all observers than images with higher level of distortions. Probably, the simplest for analysis is the test image #4. Meanwhile, there are images for which judgments have more variations than for other images. These are, in the first order, the test image #8, the artificial test image #25 as well as the highly textural images #5, 6, 13, 19.

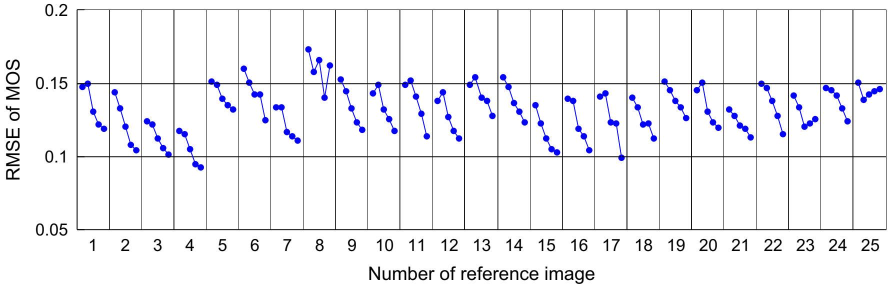


Fig. 21. Mean RMSEs of MOS for different reference images and levels of distortions.

Table 3
Distortion types and considered subsets of TID2013.

No.	Type of distortion	Noise	Actual	Simple	Exotic	New	Color	Full
1	Additive Gaussian noise	+	+	+	−	−	−	+
2	Noise in color comp.	+	−	−	−	−	+	+
3	Spatially correl. 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 transm. Errors	−	−	−	+	−	−	+
13	JPEG2000 transm. errors	−	−	−	+	−	−	+
14	Non ecc. patt. Noise	−	−	−	+	−	−	+
15	Local block-wise dist.	−	−	−	+	−	−	+
16	Mean shift	−	−	−	+	−	−	+
17	Contrast change	−	−	-	+	−	−	+
18	Change of color saturation	−	−	−	−	+	+	+
19	Multipl. Gauss. Noise	+	+	−	−	+	−	+
20	Comfort noise	−	−	−	+	+	−	+
21	Lossy compr. of noisy images	+	+	−	−	+	−	+
22	Image color quant. w. dither	−	−	−	−	+	+	+
23	Chromatic aberrations	−	−	−	+	+	+	+
24	Sparse sampl. and reconstr.	−	−	−	+	+	−	+

5. Comparative analysis for quality metrics

Practice of analysis for visual quality metric using TID2008 has demonstrated that it is reasonable to study MOS estimated for all types of distortions as well as for particular subsets [11,16,43]. That is why we have used this approach for analyzing data for TID2013.

A subset is usually formed by researchers depending upon an application and it may include one or several types of distortions. Table 3 shows subsets used below for verification of quality metrics (distortions that belong to a given subset are marked by +).

The subset "Noise" contains different types of noise and distortions in conventional image processing; the subset "Actual" relates to types of distortions most common in practice of image/video processing including compression, the sunset "Simple" includes only three standard types of distortions; the subset "Exotic" corresponds to distortions that happen not often but are among the "most difficult" for visual quality metrics.

In addition to these subsets studied earlier, we consider also the subset "New" that includes all seven new types of distortions introduced to TID2013. The subset "Color" relates to distortion types that are in one or another manner connected with changes of color content. The column "Full" contains all 24 types of distortions and metrics that provide good results for this set can be considered universal.

Correspondence to HVS has been evaluated for the following metrics (quality indices): SFF [44], componentwise FSIM and its color version FSIMc [20], PSNR-HA and PSNR-HMA [43], SR-SIM [45], MSSIM [46], MAD index [27], IW-SSIM [19], MSDDM [47], IW-PSNR [19], color version of PSNR which takes into account color in a manner similar to PSNR-HA [43], VSNR [48], PSNR-HVS [49], PSNR-HVS-M [40], SSIM [9], NQM [50], DCTune [51], VIF and a pixel based version of VIF (VIFP) [52], UQI [53], WSNR [54], CW-SSIM [55], XYZ [56], LINLAB [57], IFC [58], BMMF [59]. Many of these metrics have been calculated using Metrix MUX Visual Quality Assessment Package [60].

Table 4 presents the values of Spearman rank order correlation coefficients (SROCC) for the considered metrics and the aforementioned subsets. Similarly, Table 5 contains the corresponding values of Kendall rank order correlation coefficients (KROCC) [61]. SROCC and KROCC

Table 4
SROCC values for the considered metrics for the database TID2013.

No.	Metric	Noise	Actual	Simple	Exotic	New	Color	Full	TID2008
1	PSNR	0.8217	0.8246	0.9134	0.5968	0.6190	0.5387	0.6395	0.553
2	PSNRc	0.7691	0.8026	0.8759	0.5621	0.7772	0.7344	0.6869	0.525
3	MSSIM	0.8733	0.8871	0.9053	0.8413	0.6314	0.5663	0.7872	0.853
4	SSIM	0.7574	0.7877	0.8371	0.6320	0.5801	0.5057	0.6370	0.645
5	VSNR	0.8691	0.8817	0.9121	0.7064	0.5888	0.5122	0.6809	0.707
6	VIFP	0.7835	0.8151	0.8975	0.5574	0.5921	0.5094	0.6084	0.655
7	VIF	0.8420	0.8589	0.9321	0.6282	0.5930	0.5210	0.6770	0.750
8	NQM	0.8362	0.8572	0.8752	0.5891	0.6258	0.5418	0.6349	0.624
9	WSNR	0.8804	0.8966	0.9335	0.4227	0.6471	0.5588	0.5796	0.488
10	PSNR-HVS-M	0.9061	0.9175	0.9379	0.5644	0.6474	0.5572	0.6246	0.559
11	PSNR-HVS	0.9172	0.9257	0.9507	0.6006	0.6471	0.5589	0.6536	0.594
12	PSNR-HMA	0.9147	0.9337	0.9373	0.8139	0.7382	0.6745	0.8128	0.846
13	PSNR-HA	0.9227	0.9384	0.9527	0.8247	0.7008	0.6323	0.8187	0.868
14	FSIM	0.8969	0.9108	0.9485	0.8436	0.6494	0.5650	0.8007	0.882
15	FSIMc	0.9022	0.9149	0.9472	0.8407	0.7878	0.7752	0.8510	0.884
16	SFF	0.8787	0.9058	0.9495	0.8205	0.8502	0.8316	0.8513	0.877
17	UQI	0.6482	0.6904	0.7575	0.5313	0.4935	0.4440	0.5444	0.600
18	MSDDM	0.8740	0.8877	0.9112	0.7831	0.6341	0.5456	0.7694	0.805
19	SR_SIM	0.9070	0.9211	0.9547	0.8555	0.6510	0.5611	0.8070	0.891
20	DCTUNE	0.8827	0.8930	0.9096	0.4673	0.8443	0.8499	0.6198	0.476
21	CW_SSIM	0.7869	0.8101	0.8447	0.3859	0.6356	0.6320	0.5616	0.478
22	IFC	0.7218	0.7608	0.7792	0.3610	0.5444	0.4449	0.5400	0.569
23	IWPSNR	0.8961	0.9097	0.9237	0.6510	0.6470	0.5533	0.6888	0.682
24	IWSSIM	0.8783	0.8934	0.9173	0.8367	0.6287	0.5582	0.7774	0.856
25	Linlab	0.8577	0.8701	0.8990	0.4374	0.8535	0.8432	0.6495	0.487
26	MAD_index	0.8899	0.9032	0.9243	0.8006	0.6490	0.5623	0.7807	0.834
27	XYZ	0.8666	0.8625	0.8616	0.5166	0.7473	0.8062	0.6872	0.577
28	BMMF	0.9430	0.9490	0.9520	0.8450	0.6870	0.6660	0.8340	0.947

Table 5
KROCC values for the considered metrics for TID2013.

№	Metric	Noise	Actual	Simple	Exotic	New	Color	Full	TID2008
1	PSNR	0.6236	0.6242	0.7452	0.4254	0.4728	0.4156	0.4700	0.402
2	PSNRc	0.5619	0.5961	0.6892	0.3923	0.5761	0.5359	0.4958	0.369
3	MSSIM	0.6802	0.6982	0.7210	0.6477	0.4952	0.4557	0.6079	0.660
4	SSIM	0.5515	0.5768	0.6286	0.4548	0.4226	0.3823	0.4636	0.468
5	VSNR	0.6761	0.6908	0.7312	0.5193	0.4374	0.3788	0.5077	0.536
6	VIFP	0.5873	0.6217	0.7143	0.4066	0.4512	0.3930	0.4567	0.495
7	VIF	0.6590	0.6729	0.7694	0.4634	0.4474	0.3998	0.5148	0.586
8	NQM	0.6413	0.6665	0.6812	0.4120	0.4831	0.4087	0.4662	0.461
9	WSNR	0.6963	0.7186	0.7728	0.2973	0.5150	0.4363	0.4463	0.393
10	PSNRHVSM	0.7331	0.7495	0.7801	0.4032	0.5179	0.4409	0.4818	0.449
11	PSNRHVS	0.7547	0.7661	0.8092	0.4356	0.5169	0.4486	0.5077	0.476
12	PSNRHMA	0.7448	0.7775	0.7853	0.6101	0.5723	0.5073	0.6316	0.654
13	PSNRHA	0.7603	0.7874	0.8182	0.6245	0.5416	0.4776	0.6433	0.689
14	FSIM	0.7160	0.7371	0.7952	0.6555	0.5236	0.4524	0.6300	0.698
15	FSIMc	0.7231	0.7427	0.7929	0.6519	0.6120	0.5925	0.6669	0.699
16	SFF	0.6915	0.7316	0.8034	0.6179	0.6597	0.6347	0.6588	0.688
17	UQI	0.4601	0.4976	0.5499	0.3776	0.3529	0.3154	0.3906	0.435
18	MSDDM	0.6862	0.6978	0.7299	0.6072	0.4906	0.4237	0.5954	0.616
19	SR_SIM	0.7342	0.7563	0.8118	0.6759	0.5271	0.4489	0.6417	0.715
20	DCTUNE	0.7017	0.7167	0.7389	0.3168	0.6475	0.6488	0.4704	0.372
21	CW_SSIM	0.6128	0.6409	0.6925	0.2733	0.4884	0.4851	0.4196	0.349
22	IFC	0.5273	0.5630	0.5740	0.2579	0.3982	0.3209	0.3959	0.381
23	IWPSNR	0.7240	0.7465	0.7705	0.4606	0.5185	0.4349	0.5250	0.524
24	IWSSIM	0.6894	0.7110	0.7414	0.6414	0.4919	0.4411	0.5998	0.665
25	Linlab	0.6761	0.6942	0.7462	0.3055	0.6617	0.6483	0.4946	0.381
26	MAD_index	0.7029	0.7256	0.7519	0.6045	0.5183	0.4384	0.6035	0.645
27	XYZ	0.6746	0.6755	0.6828	0.3679	0.5371	0.6045	0.5110	0.434
28	BMMF	0.7920	0.8030	0.8070	0.6400	0.5260	0.5160	0.6640	–

values for the full set of the database TID2008 are given in the rightmost column of Tables 4 and 5, respectively, for comparison purposes.

We prefer to analyze rank order correlation coefficients between MOS and quality metrics because this allows avoiding fitting procedures that can be not unique. In fact, we have also analyzed Pearson correlation coefficients for data fitting by means of third order polynomials. The results obtained are in good agreement with the data for SROCC and KROCC. Thus, only SROCC and KROCC data are represented and analyzed below.

The first conclusion that follows from analysis of data presented in Table 4 is that even the best metrics (SFF and FSIMc) provide SROCC about 0.85 for all types of distortions (see data in column Full) and it is worse than the best metrics for the set Full of TID2008 (over 0.9, see data in the rightmost column of Table 4). This shows that the database TID2013 is really challenging for HVS-metrics and we have gained one of our intentions.

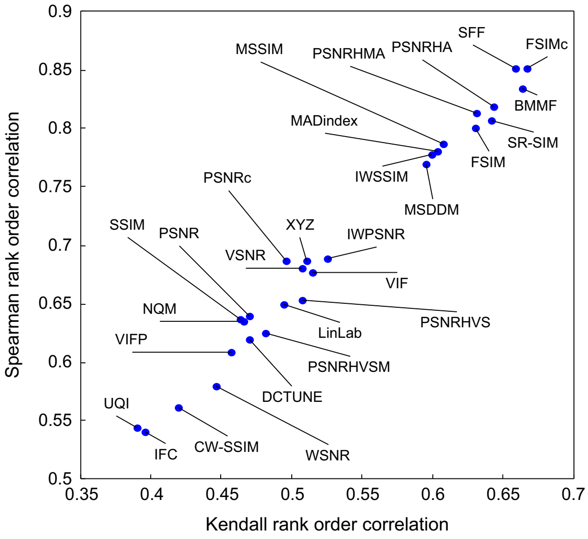


Fig. 22. SROCC and KROCC values for the considered metrics.

Consider now particular subsets. For the subset "Noise", the situation is rather good since there are several metrics (BMMF, PSNR-HA, PSNR-HVS, PSNR-HMA, PSNR-HVS-M, FSIMc, SR_SIM) for which SROCC is larger than 0.9, i.e. appropriate adequateness is provided. For the subset "actual", the situation is similar. There are quite many metrics that provide SROCC over 0.9 and reaching almost 0.95 for the best metrics. The situation is even better for the subset "Simple" where even the standard PSNR possesses SROCC over 0.9 with MOS and where the best visual quality metrics possess SROCC values over 0.95.

In turn, the subset "Exotic" causes problems for many metrics. Only a few metrics have SROCC with MOS about 0.85 (SR_SIM, FSIM, MSSIM, BMMF), i.e. the task is still not fully solved. Similarly, the situation is problematic for the subset "New". There are only three metrics that provide SROCC about 0.85 (LINLAB, DCTune, and SFF), for other metrics the SROCC values are less than 0.8. This means that, to be good enough, other metrics have to be modified and adapted to new types of distortions. The situation with the subset "Color" is at the moment not optimistic too. The best metrics provide SROCC about 0.85 and these metrics are LINLAB, DCTune and SFF. Note that all these metrics are intended just for assessment of color image visual quality.

Table 6
SROCC and KROCC values for three groups of distorted images.

Metric	Spearmen correlation			Kendall correlation		
MOS	Bad quality	Middle quality	Good quality	Bad quality	Middle quality	Good quality
FSIM	**0.7292**	0.4377	0.1977	**0.5317**	0.3074	0.1388
FSIMc	**0.7269**	**0.4776**	0.2173	**0.5280**	**0.3326**	0.1516
MSSIM	0.6581	0.4094	0.2109	0.4715	0.2854	0.1478
NQM	0.4760	0.2808	−0.0125	0.3374	0.1914	−0.0061
PSNR	0.5381	0.2642	0.1246	0.3742	0.1827	0.0909
PSNRc	0.4462	0.3017	0.1583	0.3052	0.2056	0.1107
PSNR-HA	**0.7184**	0.4426	0.2844	**0.5308**	**0.3151**	0.1959
PSVR-HMA	0.6963	**0.4547**	0.1850	0.5151	**0.3210**	0.1255
PSNR-HVS	0.6764	0.3523	0.0495	0.4981	0.2525	0.0427
PSNR-HVS-M	0.6438	0.3479	0.0139	0.4709	0.2492	−0.0048
SSIM	0.4476	0.2195	**0.3550**	0.3030	0.1522	**0.2433**
VIFP	0.6414	0.1637	**0.3694**	0.4584	0.1115	**0.2574**
VSNR	0.5245	0.3290	0.0365	0.3689	0.2268	0.0281
WSNR	0.5320	0.3203	−0.0045	0.3837	0.2220	0.0029
uqi	0.5175	0.1313	0.2143	0.3621	0.0906	0.1455
Sff (c)	0.6891	**0.4434**	0.2355	0.5008	0.3067	0.1604
dctune	0.3835	0.4080	0.0178	0.2700	0.2837	0.0203
Sr_sim (g)	**0.7527**	**0.4515**	0.2105	**0.5588**	**0.3198**	0.1492
msddm	0.6326	0.4287	0.1603	0.4438	0.3006	0.1161
iwssim	0.6658	0.3973	0.1745	0.4809	0.2760	0.1231
iwpsnr	0.5254	0.3508	0.1144	0.3905	0.2436	0.0848
Mad index	0.6419	0.4288	0.1874	0.4637	0.2995	0.1317
cwssim	0.2473	0.2579	0.0350	0.1695	0.1845	0.0313
Ifc	0.5598	0.1354	0.2579	0.3896	0.0930	0.1746
Xyz	0.4618	0.3704	0.0798	0.3225	0.2542	0.0591
VIF	0.6490	0.2341	**0.3609**	0.4655	0.1625	**0.2501**
linlab	0.4433	0.3825	0.0802	0.3146	0.2644	0.0647
BMMF	0.6431	**0.4842**	**0.4733**	0.4653	**0.3458**	**0.3263**

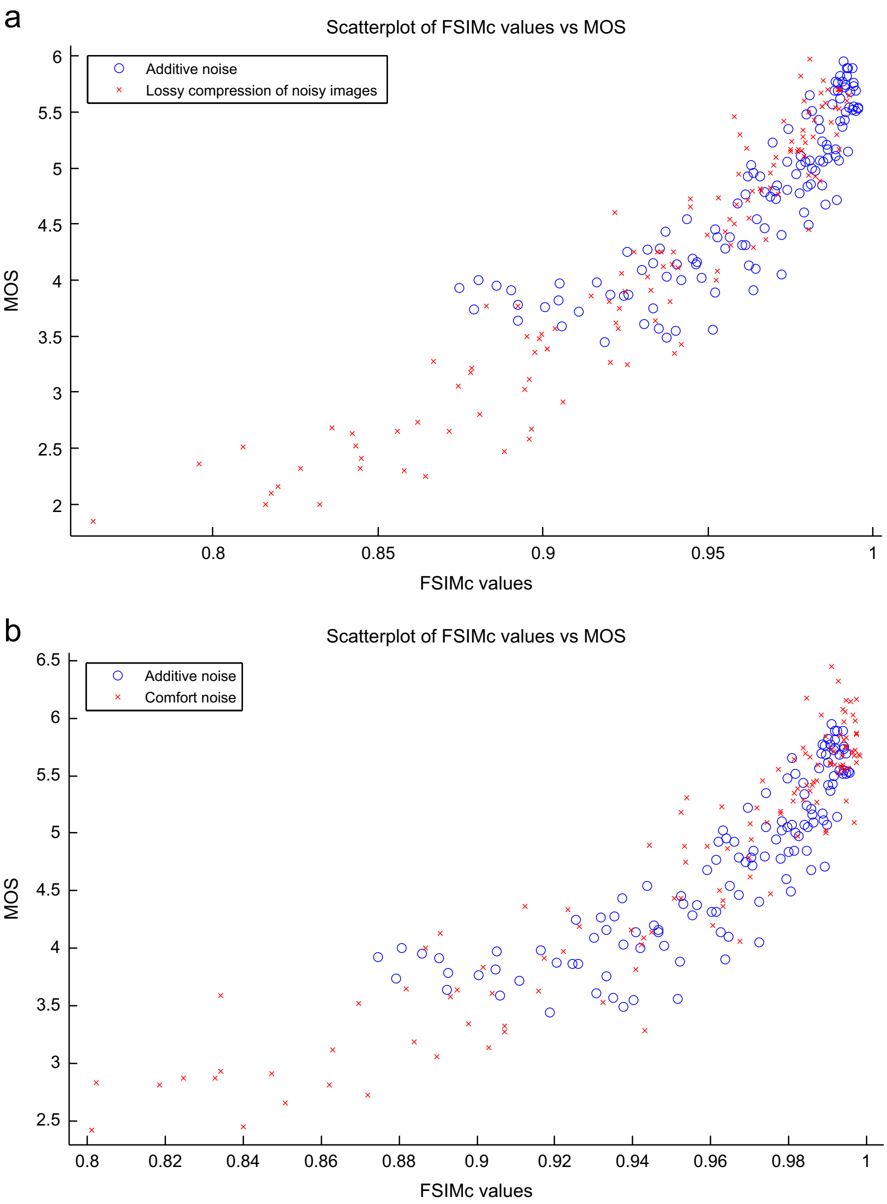


Fig. 23. Examples of scatter-plots for pairs of distortion types for which the considered metric is adequate.

Analysis of data obtained for KROCC (Table 5) leads to the same conclusions. The only difference is that KROCC values are by about 15…30% smaller than the corresponding SROCC values. The metrics SFF and FSIMc produce the most "stable" results for all subsets and they are the "winners" at the moment. To our opinion, there are three features of FSIMc that jointly provide the best results for this metric. They are good properties of SSIM put into FSIMc basis, ability to take into account color, and local adaptivity, i.e. paying more "attention" to locally active areas as details, edges, etc.

Joint analysis of the results for both rank coefficients can be performed conveniently using representation in Fig. 22. Here horizontal and vertical axes correspond to SROCC and KROCC, respectively. The best are those metrics the points for which are closer to the upper right corner. Positions of the points in this representation show that there is almost linear dependence between SROCC and KROCC that allows analyzing only one of these coefficients, e.g., SROCC.

We have also carried out specific analysis that, to the best of our knowledge, has not been done earlier. Let us divide the 3000 distorted images into three groups according to MOS obtained for them in experiments. Each group contains 1000 images and the first one is called "Bad quality" with MOS values from 0.242 to 3.94. The second group called "Middle quality" includes images with MOS from 3.94 to 5.25. Finally, the third group contains "Good quality" images with MOS larger than 5.25.

Then, let us calculate SROCC and KROCC between the considered metrics and MOS separately for each group.

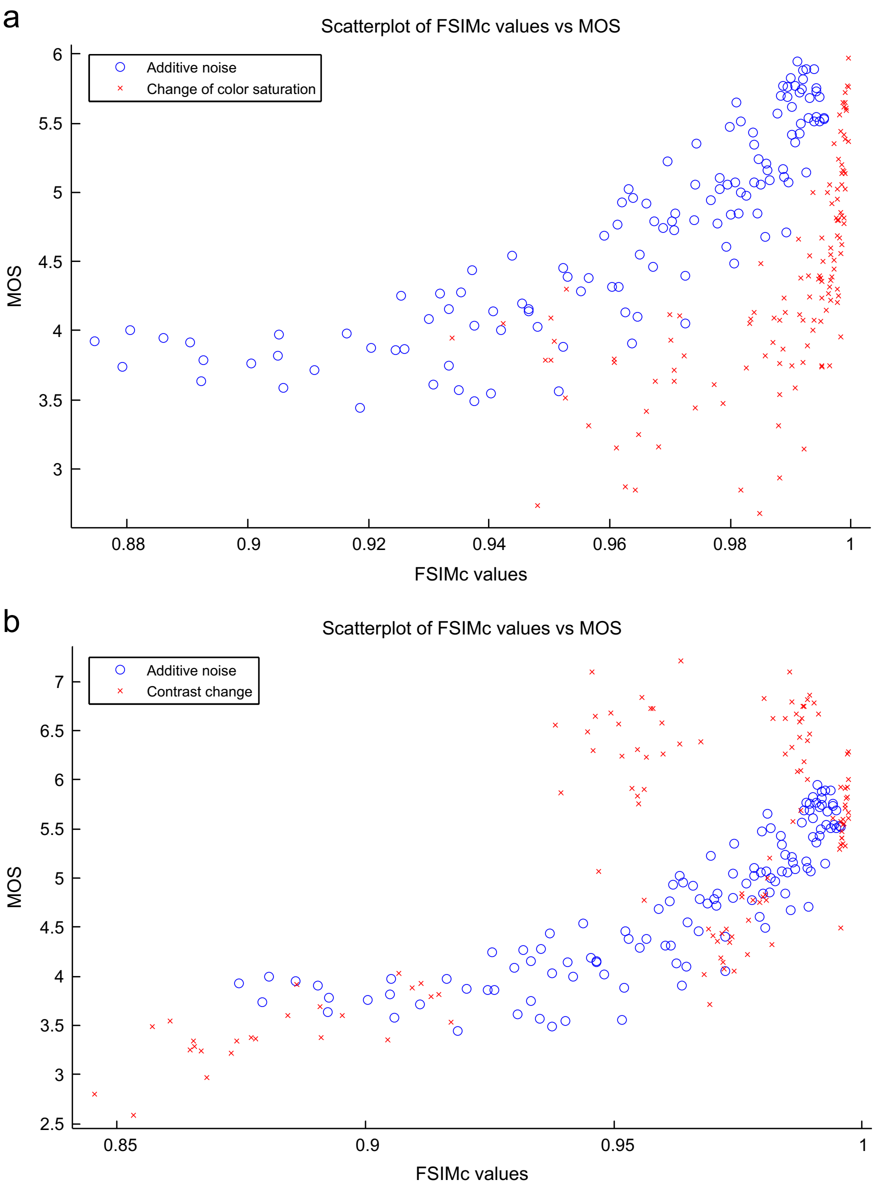


Fig. 24. Examples of scatter-plots for pairs of distortion types for which the considered metric is not adequate.

The obtained values are presented in Table 6. The results might seem quite surprising. Only for the group “Bad quality” rank correlations are high enough and the leaders are the metrics SR_SIM, FSIM, FSIMc and PSNR-HA.

For the group “Middle quality”, rank correlations are considerably smaller (not larger than 0.48 for SROCC and 0.34 for KROCC). The leaders are the same and the metrics SFF and BMMF join them. Meanwhile, for “Good quality” image group, SROCC and KROCC values are small (even negative for some metrics). The best results are provided by the metric BMMF which is an obvious leader. Next positions are occupied by the metrics SSIM, VIF and VIFP.

To our opinion, such results can be attributed to several factors. First, there is high density of MOS (see its histogram, in Fig. 17) for “Good quality” image group. For many images of this group, distortions are practically invisible and then experiment participants, having met with two such images presented at the monitor, select the best quality image in a random manner. Second, for “Bad quality” image group, experiment participants pay less attention to visual quality than to possibility to understand what is presented at picture (to information recovery from severely corrupted data [27]). Thus, this ability of humans is to be studied specially.

6. Methodology of metric drawback detection

As it can be noticed, there is no universal metric that can be considered appropriate. Therefore, a task of further studies could be detection of drawbacks (difficult distortion types) for visual quality metrics with the aim to improve their performance. Below we describe one possible way to detect such distortion types. Examples are given for the metric FSIMc which is among the best and most stable according to the results of analysis presented above.

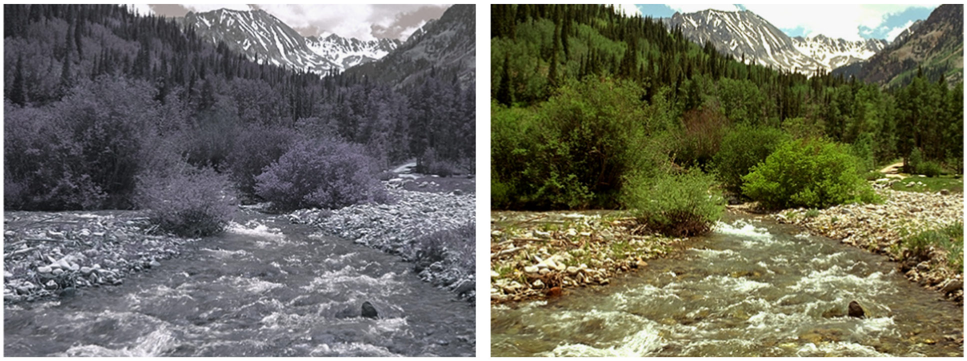
Recall that for a good metric the scatterplot of MOS and metric values behaves as it is shown in Fig. 23a. Data for two distortion types are presented here, additive white noise (#1) and lossy compression of noisy images (#21). 'Additive white noise' is chosen as a type of distortion most studied earlier for which most visual quality metrics behave properly (their values become worse if noise variance increases).

An obvious tendency to MOS increasing with increase of the metric values is observed, the points are clustered well along imaginary line fitted into scatter-plot and the points for both types of distortions are clustered together. This means that the metric is able to adequately characterize visual quality of images corrupted by both considered types of distortions (see Fig. 23b).

Consider now two other pairs of distortion types. The first pair is 'Additive white noise' and 'Change of color saturation' (#18)—see the scatterplot in Fig. 24a. It is seen that points for different distortion types belong to separate clusters that only partly intersect. The metric FSIMc over-estimates visual quality of images corrupted by chromatic aberrations assigning quite large values to them although observers do not highly assess their visual quality. An opposite case takes place for 'Additive white noise' and 'Contrast change' (#17)—see the scatter-plot in Fig. 24b. There are two obvious clusters for 'Contrast change' distortion type that are outside the "mainstream". These two clusters, in fact, correspond to images with increased contrast for which the metric FSIMc underestimates their quality.

Therefore, pairwise analysis of scatter-plots allows detecting such distortion type(s) for which a studied metric is not adapted well. Below we present several examples of situations when FSIMc value votes in favor of better visual quality of a certain image in an analyzed pair of distorted images although MOS evidences the opposite. A first example is given in Fig. 25. In this figure, image numbers in TID2013, FSIMc values and MOS are given under images. As it is seen, FSIMc is slightly larger for the image placed left (distorted by 'Change of color saturation') although MOS and visual appearance are obviously better for the image placed right (distorted by 'Contrast change').

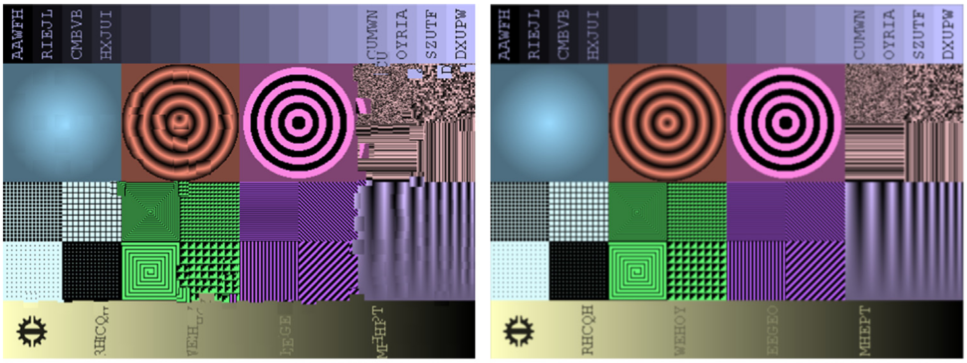
Another example is given in Fig. 26. Again FSIMc is slightly larger for the image placed left (distorted by 'Non eccentricity' pattern noise). These distortions strike the eye and, because of this, MOS for this image is considerably smaller than for the image placed right distorted by 'Gaussian blur' of level 2. A third example is given in Fig. 27 for the test image #3. The image placed left has larger



i13_18_5, MOS=3.8, FSIMc=0.948

i13_17_4, MOS=6.5, FSIMc=0.945

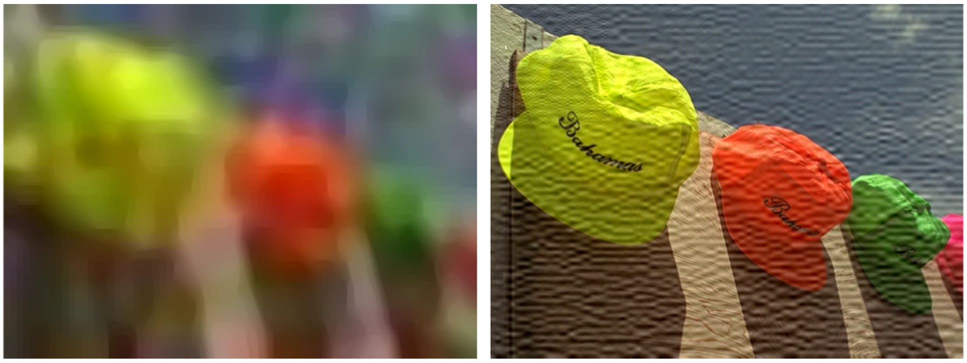
Fig. 25. Example of contradiction between FSIMc and MOS for the test image # 13.



i25_14_5, MOS=3.1, FSIMc=0.959

i25_08_2, MOS=5.1, FSIMc=0.957

Fig. 26. Example of contradiction between FSIMc and MOS for the test image # 25.



i03_24_5, MOS=0.4, FSIMc=0.693

i03_13_4, MOS=2.9, FSIMc=0.671

Fig. 27. Example of contradiction between FSIMc and MOS for the test image # 25.

FSIMc (this is the case of distortions due to compressive sensing). The image placed right obviously has better visual quality (it is corrupted by 'JPEG2000' transmission errors) and this is confirmed by considerably larger MOS.

These examples demonstrate that even the best among existing visual quality metrics are not perfect. And this stimulates further research.

7. Access to TID2013, conclusions and acknowledgements

The archive TID2013 is available for free downloading from http://ponomarenko.info/tid2013.htm. The archive includes image files, the file containing the MOS values, the file containing the RMSE of MOS, the programs for calculation of Spearman and Kendall correlations, the Readme file that explains how to exploit the database. Also, archive contains the values of most known quality metrics calculated for TID2013. Note that TID2013 occupies about 1.7 GB on a hard disk and about 900 MB in the archive.

We would like to underline the following advantages of TID2013. It contains many different types of distortion that deal with various peculiarities of HVS. Seven new types of distortions and one new level of distortions have been added to TID2013 compared to TID2008. The created database is not simple for existing visual quality metrics. One approach to analyze types of distortions difficult for a given metric is described.

The authors would like to thank all the people in Finland, Ukraine, France, Italy and USA who assisted in the experiments performance.

References

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

[2] H.R. Wu, W. Lin, L.J. Karam, An overview of perceptual processing for digital pictures, in: IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Melbourne, Australia, 2012, pp. 113–120.

[3] W. Lin, C.C. Jay Kuo, Perceptual visual quality metrics: a survey, J. Visual Commun. Image Represent. 22 (2011) 297–312.

[4] A.K. Moorthy, A.C. Bovik, Visual quality assessment algorithms: what does the future hold? Multimedia Tools Appl 51 (2011) 675–696.

[5] D.M. Chandler, Seven challenges in image quality assessment: past, present, and future research, ISRN Signal Process. 2913 (2013) 1–53.

[6] Digital Images and Human Vision, in: A.B. Watson (Ed.), MIT Press, London, England, 1993, pp. 139–140.

[7] M. Carli, Perceptual Aspects in Data Hiding, Thesis for the Degree of Doctor of Technology, Tampere University of Technology, Tampere, Finland, 2008.

[8] D.V. Fevralev, N.N. Ponomarenko, V.V. Lukin, S.K. Abramov, K.O. Egiazarian, J.T. Astola, Efficiency analysis of DCT-based filters for color image database, in: Image Processing: Algorithms and Systems IX, San Francisco, USA, 2011, 78700R-78700R-78712.

[9] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.

[10] 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 March 2014.

[11] 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), TID2008 ⟨http://ponomarenko.info/tid2008.htm⟩, 27 March 2014, pp. 30–45.

[12] Y. Horita, K. Shibata, Z.M. Parvez Saddad, Subjective Quality Assessment Toyama Database, in ⟨http://mict.eng.u-toyama.ac.jp/mict/⟩, 27 March 2014.

[13] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Electron. Imaging 19 (2010) (2014) 1–21. (CSIQ page).

[14] S. Li, L. Ma, K.N. Ngan, Full-reference video quality assessment by decoupling detail losses and additive impairments, IEEE Trans. Circuits Syst. Video Technol. 2(2012) 1100–1112, IVPL Database: ⟨http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.shtm⟩ l, 7 July 2014, no. 99, 2012.

[15] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.C.J. Kuo, Color image database TID2013: peculiarities and preliminary results, in: Fourth European Workshop on Visual Information Processing (EUVIP), 2013, pp. 106–111.

[16] N. Ponomarenko, O. Ieremeiev, V. Lukin, L. Jin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.C.J. Kuo, A new color image database TID2013: innovations and results, in: J. Blanc-Talon, A. Kasinski, W. Philips, D. Popescu, P. Scheunders (Eds.), Advanced Concepts for Intelligent Vision Systems, Springer International Publishing, 2013, pp. 402–413.

[17] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, F. Battisti, Color image database for evaluation of image quality metrics, in: IEEE 10th Workshop on Multimedia Signal Processing, 2008, pp. 403–408.

[18] S. Winkler, Analysis of public image and video databases for quality assessment, IEEE J. Sel. Top. Signal Process. 6 (2012) 616–625.

[19] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Trans. Image Process. 20 (2011) 1185–1198.

[20] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (2011) 2378–2386.

[21] L. Zhang, L. Zhang, X. Mou, D. Zhang, A comprehensive evaluation of full reference image quality assessment algorithms, in: 19th IEEE International Conference on Image Processing (ICIP), 2012, pp. 1477–1480.

[22] J. Lina, K. Egiazarian, C.C.J. Kuo, Perceptual image quality assessment using block-based multi-metric fusion (BMMF), in: IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 1145–1148.

[23] M. Uss, B. Vozel, V. Lukin, S. Abramov, I. Baryshev, K. Chehdi, Image informative maps for estimating noise standard deviation and texture parameters, EURASIP J. Adv. Signal Process. 2011 (2011) 806516.

[24] S. Pyatykh, J. Hesser, Z. Lei, Image Noise Level, Estimation by principal component analysis, IEEE Trans. Image Process. 22 (2013) 687–699.

[25] N. Ponomarenko, V. Lukin, K. Egiazarian, HVS-metric-based performance analysis of image denoising algorithms, in: Third European Workshop on Visual Information Processing (EUVIP), 2011, pp. 156–161.

[26] C.T. Vu, T.D. Phan, D.M. Chandler, A spectral and spatial measure of local perceived sharpness in natural images, IEEE Trans. Image Process. 21 (2012) 934–945.

[27] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, ELECTIM 19 (2010). (011006-011006-011021).

[28] M.R. Peres, The Focal Encyclopedia of Photography, fourth ed. Focal Press, 2007.

[29] N.N. Ponomarenko, V.V. Lukin, O.I. Ieremeyev, K.O. Egiazarian, J.T. Astola, Visual quality analysis for images degraded by different types of noise, in: Image Processing: Algorithms and Systems XI, San Francisco, USA, 2013, 86550E-86550E-86511.

[30] M. Iqbal, J. Chen, W. Yang, P. Wang, B. Sun, SAR image despeckling using selective 3D filtering of multiple compressive reconstructed images, Prog. Electromagnet. Res. 134 (2013) 209–226.

[31] R. Kumar, M. Rattan, Analysis of various quality metrics for medical image processing, Int. J. Adv. Res. Comput. Sci. Software Eng. 2 (2012) 137–144.

[32] B.T. Oh, C.C.J. Kuo, S. Sun, S. Lei, Film grain noise modeling in advanced video coding, in: Visual Communications and Image Processing, San Jose, USA, 2007, 650811-650811-650812.

[33] D. Petrescu, J. Pincenti, Quality and noise measurements in mobile phone video capture, in: Multimedia on Mobile Devices 2011; and Multimedia Content Access: Algorithms and Systems V, San Francisco, USA, 2011, 788105-788105-788114.

[34] N.N. Ponomarenko, V.V. Lukin, K.O. Egiazarian, L. Lepisto, Adaptive visually lossless JPEG-based color image compression, SIViP 7 (2013) 437–452.

[35] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, ADCT: a new high quality DCT based coder for lossy image compression, in: Proceedings of International Workshop on Local and Non-Local Approximation in Image Processing (LNLA '08), Lausanne, Switzerland, 2008, pp. 1–6.

[36] M. Pedersen, N. Bonnier, J.Y. Hardeberg, F. Albregtsen, Attributes of image quality for color prints, J. Electron. Imaging 19 (2010) 011016.

[37] J.L. Paredes, G.R. Arce, Compressive sensing signal reconstruction by weighted median regression estimates, IEEE Trans. Signal Process. 59 (2011) 2585–2601.

[38] A. Danielyan, A. Foi, V. Katkovnik, K. Egiazarian, Spatially adaptive filtering as regularization in inverse imaging, Super-Resolution Imaging, CRC Press, 2010, 123–153.

[39] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (2006) 3440–3451.

[40] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, V. Lukin, On between-coefficient contrast masking of DCT basis functions, in: Third International Workshop on Video Processing and Quality Metrics, Scottsdale, USA, 2007, pp. 1–4.

[41] ITU, Methodology for the subjective assessment of the quality of television pictures, in: Recommendation BT.500-11, Geneva, Switzerland, 2002.

[42] N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, V. Lukin, Metrics performance comparison for color image database, in: Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, USA, 2009, pp. 1–6.

[43] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, M. Carli, Modified image visual quality metrics for contrast change and mean shift accounting, in: 11th International Conference the Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Polyana-Svalyava, Ukraine, 2011, pp. 305–311.

[44] H.-W. Chang, H. Yang, Y. Gan, M.-H. Wang, Sparse feature fidelity for perceptual image quality assessment, IEEE Trans. Image Process. 22 (2013) 4007–4018.

[45] L. Zhang, H. Li, SR-SIM: a fast and high performance IQA index based on spectral residual, in: 19th IEEE International Conference on Image Processing (ICIP), Orlando, USA, 2012, pp. 1473–1476.

[46] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2003, pp. 1398–1402.

[47] N. Ponomarenko, L. Jin, V. Lukin, K. Egiazarian, Self-similarity measure for assessment of image visual quality, in: J. Blanc-Talon, R. Kleihorst, W. Philips, D. Popescu, P. Scheunders (Eds.), Advanced Concepts for Intelligent Vision Systems, Springer, Berlin Heidelberg, 2011, pp. 459–470.

[48] D.M. Chandler, S.S. Hemami, VSNR: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (2007) 2284–2298.

[49] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, M. Carli, New full-reference quality metrics based on HVS, in: Second International Workshop on Video Processing and Quality Metrics, Scottsdale, USA, 2006, pp. 1–4.

[50] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, A.C. Bovik, Image quality assessment based on a degradation model, IEEE Trans. Image Process. 9 (2000) 636–650.

[51] A.B. Watson, DCTune: a technique for visual optimization of DCT quantization matrices for individual images, in: Society for Information Display Digest of Technical Papers XXIV, 1993, pp. 946–949.

[52] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2006) 430–444.

[53] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (2002) 81–84.

[54] T. Mitsa, K.L. Varkur, Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Minneapolis, USA, 1993, pp. 301–304.

[55] Z. Wang, E.P. Simoncelli, Translation insensitive image similarity in complex wavelet domain, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05), Philadelphia, USA, 2005, pp. 573–576.

[56] B.W. Kolpatzik, C.A. Bouman, Optimized universal color palette design for error diffusion, ELECTIM 4 (1995) 131–143.

[57] B.W. Kolpatzik, C.A. Bouman, Optimized error diffusion for image display, ELECTIM 1 (1992) 277–292.

[58] H.R. Sheikh, A.C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (2005) 2117–2128.

[59] L. Jin, S. Cho, T.-J. Liu, K. Egiazarian, C.C.J. Kuo, Performance comparison of decision fusion strategies in BMMF-based image quality assessment, in: Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Hollywood, USA, 2012, pp. 1–4.

[60] M. Gaubatz, Metrix MUX Visual Quality Assessment Package, in 〈http://foulard.ece.cornell.edu/gaubatz/metrix_mux/〉, 27 March 2014.

[61] M.G. Kendall, The Advanced Theory of Statistics, Charles Griffin & Company Limited, London, UK, 1945.